UNIVERSITY OF CALIFORNIA, SAN DIEGO

Reproducing Kernel Space Embeddings and Metrics on Probability Measures

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Bharath Kumar Sriperumbudur Vangeepuram

Committee in charge:

Gert R. G. Lanckriet, Chair
Ian Abramson
Ery Arias-Castro
Sanjoy Dasgupta
Kenneth Kreutz-Delgado
Bhaskar D. Rao
Lawrence K. Saul
Bernhard Schölkopf

2010

The dissertation of Bharath Kumar Sriperumbudur Vangeepuram is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2010

DEDICATION

To Appa and Amma for their love and support

&

To all those who believe in learning for its own sake

EPIGRAPH

Embed me in a Hilbert space and
He will solve my problems.
—Probability Measure

TABLE OF CONTENTS

Signature Page . . . iii

Dedication . . . iv

Epigraph . . . v

Table of Contents . . . vi

List of Figures . . . ix

List of Tables . . . x

Acknowledgements . . . xi

Vita . . . xiii

Abstract of the Dissertation . . . xv

1 Introduction . . . 1
- 1.1 Learning and Inference in Reproducing Kernel Hilbert Spaces . . . 2
- 1.2 Summary of Contributions . . . 5

2 Kernels and Reproducing Kernel Hilbert Spaces . . . 9
- 2.1 Positive Definite Functions and Kernels . . . 9
- 2.2 Integrally and Conditionally Positive Definite Functions . . . 14
- 2.3 The Reproducing Kernel Hilbert Space of a Kernel . . . 16

3 Characteristic Kernels and Maximum Mean Discrepancy . . . 18
- 3.1 Interpretation of MMD . . . 19
- 3.2 Characteristic Kernels . . . 25
  - 3.2.1 Integrally Strictly Positive Definite Kernels are Characteristic 28
  - 3.2.2 Characterization for Translation Invariant Kernels on $\mathbb{R}^d$ . . 30
  - 3.2.3 Characterization for Translation Invariant Kernels on $\mathbb{T}^d$ . . 34
  - 3.2.4 Proofs . . . 36
- 3.3 Dissimilar Distributions with Small MMD . . . 44
- 3.4 Discussion . . . 49
- Bibliographic Notes . . . 50

4 Universality, Characteristic Kernels and Other Notions . . . 52
- 4.1 Universal Kernels . . . 53
  - 4.1.1 *c*-universality . . . 54
  - 4.1.2 *cc*-universality . . . 54

4.1.3 c_0- and L_p-universality 56
4.1.4 Summary and Open Questions 57
4.2 Characteristic vs. Universal Kernels 58
4.3 Universal & Characteristic Kernels vs. Others 62
4.4 New Results 65
4.4.1 c_0-universality and RKHS Embedding of Measures 65
4.4.2 Universal vs. Integrally Strictly Positive Definite Kernels . . 67
4.4.3 Radial kernels on $\mathbb{R}^d$ 67
4.4.4 Characteristic vs. Conditionally Strictly pd Kernels 68
4.5 c_b-universality 71
Bibliographic Notes 72

5 Integral Probability Metrics, ϕ-Divergences and MMD 74
5.1 Introduction 74
5.2 Empirical Estimation of IPMs 79
5.2.1 Empirical Estimators of W, β and γ_k 80
5.2.2 Interpretability of IPMs: Relation to Binary Classification . 84
5.2.3 Consistency and Rate of Convergence 88
5.2.4 Simulation Results 95
5.2.5 Empirical Estimation of Total Variation Distance 103
5.3 Metrization of the Weak Topology 107
5.4 Discussion 113
Bibliographic Notes 113

6 Choice of Characteristic Kernel and Two-Sample Test 114
6.1 Generalizing the MMD for Classes of Characteristic Kernels 115
6.2 Estimation of γ: Consistency and Rate of Convergence 117
6.3 Experiments 122
Bibliographic Notes 124

7 Banach Space Embedding of Probability Measures 126
7.1 Preliminaries: Reproducing Kernel Banach Spaces 128
7.2 RKBS Embedding of Probability Measures 131
7.2.1 When is (7.6) Injective? 133
7.2.2 Consistency Analysis 138
7.2.3 Computation of $\gamma_K(\mathbb{P},\mathbb{Q})$ 140
7.3 Concrete Examples of RKBS Embeddings 141
7.4 Discussion 150
Bibliographic Notes 151

A Relation Between IPMs and ϕ-Divergences 152

B Derivation of (7.23)–(7.25) . . . 155
B.1 Derivation of (7.23) . . . 155
B.2 Derivation of (7.24) and (7.25) . . . 156

C Appendix . . . 158
C.1 Definitions . . . 158
C.1.1 Standard Spaces . . . 158
C.1.2 Distributions and Fourier Transforms . . . 160
C.2 Supplementary Results . . . 161
C.2.1 Real Analysis . . . 162
C.2.2 Fourier Analysis . . . 163
C.2.3 Convex Analysis . . . 163
C.2.4 Concentration Inequalities and Empirical Processes . . . 163

Bibliography . . . 165

Notation and Symbols . . . 173

Abbreviations . . . 176

LIST OF FIGURES

Figure 3.1: Illustration of Example 3.23 40
Figure 3.2: Illustration of Example 3.24 43
Figure 3.3: Illustration of Example 3.25 46
Figure 3.4: Behavior of the empirical estimate of $\gamma_k^2(\mathbb{P},\mathbb{Q})$ 47

Figure 4.1: Summary of the relations between various families of kernels . . 73

Figure 5.1: Empirical estimates of the Kantorovich metric (Example 5.14) . 97
Figure 5.2: Empirical estimates of the Kantorovich metric (Example 5.15) . 98
Figure 5.3: Empirical estimates of MMD (Example 5.16) 100
Figure 5.4: Empirical estimates of MMD (Example 5.17) 102
Figure 5.5: Empirical estimates of the Dudley metric (Example 5.18) 103

Figure 6.1: Performance comparison of γ and γ_k in a two-sample test . . . 125

LIST OF TABLES

Table 3.1: Summary of results: When is k characteristic? 51

ACKNOWLEDGEMENTS

I would like to express my gratitude to everyone who has, directly or indirectly, helped me through my time in graduate school. First and foremost, I must thank my advisor, Gert Lanckriet, who gave me intellectual freedom to pursue my research ideas, supported me financially, guided me, gave me practical advice and helped me in forging research collaborations through visits to Max Planck Institute for Biological Cybernetics (MPI), Tübingen and The Institute for Statistical Mathematics (ISM), Tokyo. My special thanks to Alon Orlitsky and Bhaskar Rao for supporting my application to attend the doctoral program at UCSD. I would like to thank my entire committee for their valuable comments on my research and special thanks to Ery Arias-Castro for creating opportunities for collaboration. It is a pleasure to thank Ery Arias-Castro, Sanjoy Dasgupta and Lawrence Saul for their career advice. Thanks to theory enthusiasts, Daniel Hsu, Samory Kpotufe and Nakul Verma for many interesting discussions and to Omer Lang and David Torres for helping me with experiments in our joint works.

Further afield, I thank Bernhard Schölkopf for supporting my application to attend the machine learning summer school at MPI and funding my internships at MPI that cumulatively lasted for nine months. It is during this internship, over one of the usual late night discussions with Arthur Gretton, that the initial ideas for this dissertation were born. Many thanks to Arthur Gretton for the countless brainstorming sessions, for being the only guy at MPI who matched my working time zones and for being my swimming coach. I profusely thank Kenji Fukumizu for teaching me mathematics and hosting me at ISM for a period of two months. I have no doubts that without Kenji, this thesis would not be as rigorous as it is now. I thank all my research collaborators, Arthur, Bernhard, Kenji and my advisor Gert for their time and patience in commenting on my publications.

Apart from research, my graduate life was enriched by interacting with many interesting and smart people, whom I thank here. This includes Aravind and Vikas for countless discussions on anything and everything for innumerous hours, Prem and Yatharth for sharing my passion for cricket, Omer and Rupert for being my squash buddies, and at large the Indian community which made my stay pleasant and enjoyable at UCSD.

Lastly, I am grateful to my parents for their love, patience and kind support, to my brother, Ranga and sister, Hari, without whom I would not be what I am, and to my wife Aishwarya, who patiently waited over the last one year for the completion of this dissertation.

Chapters 3, 5 and 6 are based on joint work with Kenji Fukumizu, Arthur Gretton, Gert Lanckriet and Bernhard Schölkopf, which appeared in [75,76,78,79]. Chapter 4 is based on joint work with Kenji Fukumizu and Gert Lanckriet, which appeared in [77]. The longer version of [77] is currently under submission to the Journal of Machine Learning Research. Chapter 7 is based on joint unpublished work with Kenji Fukumizu and Gert Lanckriet. The dissertation author was the primary investigator and author of these papers.

VITA

1999	Bachelor of Technology in Electronics and Communication Engineering, Sri Venkateswara University, Tirupati, India
2002	Master of Technology in Electrical Engineering, Indian Institute of Technology, Kanpur, India
2010	Doctor of Philosophy in Electrical Engineering (Signal and Image Processing), University of California, San Diego

PUBLICATIONS

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf and G. R. G. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, April 2010.

B. K. Sriperumbudur, D. A. Torres and G. R. G. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Machine Learning*, To appear.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf and G. R. G. Lanckriet, "Non-parametric estimation of integral probability metrics," in *Proc. of IEEE International Symposium on Information Theory*, pp. 1428–1432, June 2010.

B. K. Sriperumbudur, K. Fukumizu and G. R. G. Lanckriet, "On the relation between universality, characteristic kernels and RKHS embedding of measures," in *JMLR Workshop and Conference Proceedings*, vol. 9, pp. 781–788, AISTATS 2010.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. G. Lanckriet and B. Schölkopf, "Kernel choice and classifiability for RKHS embeddings of probability distributions," *Advances in Neural Information Processing Systems*, vol. 22, pp. 1750–1758, 2009.

B. K. Sriperumbudur and G. R. G. Lanckriet, "On the convergence of the concave-convex procedure,", *Advances in Neural Information Processing Systems*, vol. 22, pp. 1759–1767, 2009.

A. Gretton, K. Fukumizu, Z. Harchaoui and B. K. Sriperumbudur, "A fast, consistent kernel two-sample test," *Advances in Neural Information Processing Systems*, vol. 22, pp. 673–681, 2009.

K. Fukumizu, B. K. Sriperumbudur, A. Gretton and B. Schölkopf, "Characteristic kernels on groups and semigroups," *Advances in Neural Information Processing Systems*, vol. 21, pp. 473–480, 2009.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proc. of the 21st Annual Conference on Learning Theory*, pp. 111–122, 2008.

B. K. Sriperumbudur, O. Lang and G. R. G. Lanckriet, "Metric embedding for kernel classification rules," in *Proc. of the 25th International Conference on Machine Learning*, pp. 1008-1015, 2008.

B. K. Sriperumbudur, D. A. Torres and G. R. G. Lanckriet, "Sparse eigen methods by d.c. programming," in *Proc. of the 24th International Conference on Machine Learning*, pp. 831–838, 2007.

B. K. Sriperumbudur, K. Fukumizu and G. R. G. Lanckriet, "Universality, characteristic kernels and RKHS embedding of measures," *Journal of Machine Learning Research*, Submitted.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf and G. R. G. Lanckriet, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, To be submitted.

B. K. Sriperumbudur and G. R. G. Lanckriet, "A proof of convergence of the concave-convex procedure using Zangwill's theory,", *Pattern Recognition*, To be submitted.

ABSTRACT OF THE DISSERTATION

Reproducing Kernel Space Embeddings and Metrics on Probability Measures

by

Bharath Kumar Sriperumbudur Vangeepuram

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2010

Gert R. G. Lanckriet, Chair

The notion of Hilbert space embedding of probability measures has recently been used in various statistical applications like dimensionality reduction, homogeneity testing, independence testing, etc. This embedding represents any probability measure as a mean element in a reproducing kernel Hilbert space (RKHS). A pseudometric on the space of probability measures can be defined as the distance between distribution embeddings: we denote this as γ_k, indexed by the positive definite (pd) kernel function k that defines the inner product in the RKHS.

In this dissertation, various theoretical properties of γ_k and the associated RKHS embedding are presented. First, in order for γ_k to be useful in practice, it is essential that it is a metric and not just a pseudometric. Therefore, various easily checkable characterizations have been obtained for k so that γ_k is a metric (such k are referred to as *characteristic kernels*), in contrast to the previously published characterizations which are either difficult to check or may apply only in restricted circumstances (e.g., on compact domains). Second, the relation of characteristic kernels to the richness of RKHS—how well an RKHS approximates some target function space—and other common notions of pd kernels like strictly pd (spd), integrally spd, conditionally spd, etc., is studied. Third, the question of the nature of topology induced by γ_k is studied wherein it is shown that γ_k associated with integrally spd kernels—a stronger notion than a characteristic kernel—metrize the weak* (weak-star) topology on the space of probability measures. Fourth, γ_k is compared to integral probability metrics (IPMs) and ϕ-divergences, wherein it is shown that the empirical estimator of γ_k is simple to compute and exhibits fast rate of convergence compared to those of IPMs and ϕ-divergences. These properties make γ_k to be more applicable in practice than these other families of distances. Finally, a novel notion of embedding probability measures into a reproducing kernel Banach space (RKBS) is proposed and its properties are studied. It is shown that the proposed embedding and its properties generalize their RKHS counterparts, thereby resulting in richer distance measures on the space of probabilities.

1 Introduction

The concept of distance between probability measures is a fundamental one and has found applications in many areas of science and engineering, in particular in probability theory and statistics [48, 61, 62]. In statistics, distances between probability measures are used in a variety of applications, including hypothesis tests (homogeneity tests, independence tests, and goodness-of-fit tests), density estimation, Markov chain monte carlo, etc. As an example, homogeneity testing, also called the two-sample problem, involves choosing whether to accept or reject a null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ versus the alternative $H_1 : \mathbb{P} \neq \mathbb{Q}$, using random samples $\{X_j^{(1)}\}_{j=1}^m$ and $\{X_j^{(2)}\}_{j=1}^n$ drawn i.i.d. from probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a topological space $(\mathcal{X}, \mathscr{A})$. It is easy to see that solving this problem is equivalent to testing $H_0 : D(\mathbb{P}, \mathbb{Q}) = 0$ versus $H_1 : D(\mathbb{P}, \mathbb{Q}) > 0$, where D is a metric on the space of all probability measures defined on $\mathcal{X}$. The problems of testing independence (whether or not a joint probability distribution factorizes into two marginal distributions) and goodness-of-fit (whether or not a given probability measure belongs to some pre-defined family of measures—for example, whether a given probability measure is Gaussian or not) can be posed in an analogous form. In non-parametric density estimation, $D(p_n, p_0)$ can be used to study the quality of the density estimate, p_n, that is based on the samples $\{X_j\}_{j=1}^n$ drawn i.i.d. from p_0. Popular examples for D in these statistical applications include the *Kullback-Leibler divergence*, the *total variation distance*, the *Hellinger distance* [83]—these three are specific instances of ϕ-divergence [1, 15]—the *Kolmogorov distance* [47, Section 14.2], the *Wasserstein distance* [19], etc.

In probability theory, the distance between probability measures is used in studying limit theorems, the popular example being the central limit theorem.

Another application is in metrizing the weak convergence of probability measures on a separable metric space, where the *Lévy-Prohorov distance* [23, Chapter 11] and *dual-bounded Lipschitz distance* (also called the *Dudley metric*) [23, Chapter 11] are commonly used.

In this dissertation, we study the following *pseudometric* (see footnote 2 for the definition of a pseudometric) on the space of probability measures,

$$\gamma_k(\mathbb{P},\mathbb{Q}) := D(\mathbb{P},\mathbb{Q}) = \left\Vert \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{Q}(x)\right\Vert_{\mathcal{H}}, \tag{1.1}$$

which is obtained by embedding $\mathbb{P}$ into a reproducing kernel Hilbert space (RKHS) [4], $\mathcal{H}$ as $\int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x)$, i.e.,

$$\mathbb{P} \mapsto \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x) \tag{1.2}$$

and computing the distance between the embeddings of $\mathbb{P}$ and $\mathbb{Q}$ in $\mathcal{H}$. Here, k represents the reproducing kernel of $\mathcal{H}$ and $\Vert\cdot\Vert_{\mathcal{H}}$ represents the RKHS norm. We refer the reader to Chapter 2 for background on kernels and RKHSs. The distance measure in (1.1) and the associated embedding in (1.2), which were first proposed in the statistics and probability literature [9, Chapter 4], have recently gained attention in statistical machine learning [37, 72] and have been used in various applications like dimensionality reduction [28, 29], two-sample test [37], independence tests [30, 38], density estimation [73], etc. Before we summarize the contributions of this dissertation in Section 1.2, in the following section, we briefly introduce the paradigm of learning and inference in RKHS—a popular framework in statistical machine learning—which will be helpful to understand the usefulness of the embedding in (1.2).

1.1 Learning and Inference in Reproducing Kernel Hilbert Spaces

Let us consider a binary classification problem, wherein given samples $\mathcal{D} := \{(x_j,y_j)\}_{j=1}^N$, $x_j \in \mathcal{X}$, $y_j \in \{-1,+1\}$, the goal is to learn a function, $f : \mathcal{X} \to \mathbb{R}$ such that $y_j = \text{sign}(f(x_j))$. Suppose $f(x) = \langle w, x\rangle + b$, where $\mathcal{X} \subset \mathbb{R}^d$, which means we would like to find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $\langle w, x_j \rangle + b > 0$ for all j with $y_j = +1$ and $\langle w, x_j \rangle + b < 0$ for all j with $y_j = -1$. Here $\langle \cdot, \cdot \rangle$ represents the inner product in $\mathbb{R}^d$. Since the learned function, f has to *generalize* well to unknown samples (i.e., to samples not in $\mathcal{D}$), the popular idea in machine learning is to maximize the *margin* (i.e., maximize the distance from f to the points in $\mathcal{D}$), which yields the following program,

$$\max_{w,b} \min_{j \in \{1,\dots,N\}} \frac{|\langle w, x_j \rangle + b|}{\|w\|}. \tag{1.3}$$

(1.3) can be rewritten as

$$\max_{w,b,\theta} \{\theta \, : \, |\langle w, x_j \rangle + b| \geq \theta \|w\|, \, \forall \, j\},$$

which is equivalent to

$$\max_{w,b,\theta} \{\theta \, : \, y_j(\langle w, x_j \rangle + b) \geq \theta \|w\|, \, \forall \, j\},$$

i.e.,

$$\min_{w,b} \{\|w\| \, : \, y_j(\langle w, x_j \rangle + b) \geq 1, \, \forall \, j\}. \tag{1.4}$$

Having computed (w^*, b^*) that solves (1.4), the learned function is given by $f^*(x) = \langle w^*, x \rangle + b^*$, which means given any $x \in \mathbb{R}^d$, its associated label, y can be obtained as $y = \text{sign}(f^*(x))$. (1.4) is popularly referred to as the *hard-margin support vector machine* (SVM) [5] in the machine learning literature. However, one shortcoming with this algorithm is that since f is linear in x, it is not suitable to classify samples that cannot be linearly separated, i.e., for a given $\mathcal{D}$, there does not exist $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y_j = \text{sign}(\langle w, x_j \rangle + b)$ for all j. To resolve this issue, Boser et al. [5] proposed to map the input data $(x_1, \dots, x_N)$ into a (possibly infinite-dimensional) Hilbert space, $\mathcal{H}$, by a typically non-linear map $\Phi : \mathcal{X} \to \mathcal{H}$ and then apply the algorithm in (1.4) to the mapped data set $\{(\Phi(x_j), y_j)\}_{j=1}^N$. Therefore, (1.4) reduces to

$$\min_{w \in \mathcal{H}, b \in \mathbb{R}} \left\{ \frac{1}{2} \|w\|_{\mathcal{H}}^2 \, : \, y_j(\langle w, \Phi(x_j) \rangle_{\mathcal{H}} + b) \geq 1, \, \forall \, j \right\}, \tag{1.5}$$

whose Lagrangian dual is given by

$$\min_{\{\alpha_j\}_{j=1}^N} \left\{ \frac{1}{2} \sum_{l,j=1}^N \alpha_l \alpha_j y_l y_j \langle \Phi(x_l), \Phi(x_j) \rangle_{\mathcal{H}} - \sum_{j=1}^N \alpha_j \, : \, \sum_{j=1}^N y_j \alpha_j = 0, \, \alpha_j \geq 0, \, \forall \, j \right\}. \tag{1.6}$$

Supposing (w^*, b^*) and $\{\alpha_j^*\}_{j=1}^N$ solve (1.5) and (1.6) respectively, we have

$$w^* = \sum_{j=1}^{N} y_j \alpha_j^* \Phi(x_j),$$

which implies

$$f^*(x) = \sum_{j=1}^{N} y_j \alpha_j^* \langle \Phi(x_j), \Phi(x) \rangle_{\mathcal{H}} + b^*.$$

Note that the computation of $\{\alpha_j^*\}_{j=1}^N$ and f^* depend on $\{x_j\}_{j=1}^N$ through $\langle \Phi(\cdot), \Phi(\cdot) \rangle_{\mathcal{H}}$, which means if $\mathcal{H}$ is chosen to be an RKHS with k as its reproducing kernel, then the computation of α^* and f^* depend on $\{x_j\}_{j=1}^N$ only through k as $k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$ (see Chapter 2 for details). Therefore, by simply choosing a kernel function, k—for example, the Gaussian kernel, $k(x,y) = \exp(-\sigma \|x - y\|_2^2)$, $x, y \in \mathbb{R}^d$, $\sigma > 0$—the hard-margin SVM algorithm can be extended to handle data sets that are not linearly separable—observe that $k(x,y) = \langle x, y \rangle$ yields (1.4)—which means by embedding $\mathcal{D}$ into an RKHS, $\mathcal{H}$, it is possible to construct non-linear algorithms (non-linearity is in the dependence of f^* on x) from linear ones. Since $k(x,y) = \langle k(\cdot,x), k(\cdot,y) \rangle_{\mathcal{H}}$ (see Chapter 2), we can choose $\Phi(x) = k(\cdot, x)$, which means

$$x \mapsto \Phi(x) = k(\cdot, x) = \int_{\mathcal{X}} k(\cdot, y)\, d\delta_x(y),$$

where δ_x is the Dirac measure at x. This shows that embedding x into $\mathcal{H}$ as $k(\cdot, x)$ is equivalent to embedding δ_x into $\mathcal{H}$ as $k(\cdot, x)$ through (1.2). Therefore, the embedding in (1.2) is a generalization of the idea of embedding $\mathcal{D}$ into $\mathcal{H}$.

As mentioned above, embedding $\mathcal{D}$ into $\mathcal{H}$ provides a useful way of constructing non-linear algorithms from linear ones. Since (1.2) is a generalization of embedding $\mathcal{D}$ into $\mathcal{H}$, we show below that (1.2) provides a linear method for dealing with the higher-order statistics of random variables. Let us consider the embedding in (1.2), which can rewritten as

$$\mathbb{P} \mapsto \mathbb{E}\left[k(\cdot, X)\right] = \mathbb{E}\left[\Phi(X)\right],$$

where X is a $\mathcal{X}$-valued random variable that is distributed according to $\mathbb{P}$. Here, $\mathbb{E}$ represents the expectation w.r.t. $\mathbb{P}$. It is well known that $\mathbb{P}$ can be completely characterized through its moments. The advantage with the above embedding is that by appropriately choosing Φ, $\mathbb{P}$ can be completely characterized by just computing the first moment of the embedded random variable, $\Phi(X)$. To show this, suppose

$$k(x,y)=c_0+c_1(xy)+c_2(xy)^2+\cdots,\ \ c_j\neq 0,\forall\, j\in\mathbb{N}.$$

This means

$$\mathbb{E}\left[k(y,X)\right]=c_0+c_1\mathbb{E}[X]y+c_2\mathbb{E}[X^2]y^2+\cdots,$$

which contains all the moments of $\mathbb{P}$ and therefore characterizes $\mathbb{P}$ completely. If $k(x,y)=e^{\langle x,y\rangle}$, $x,y\in\mathbb{R}^d$, we obtain the moment generating function of $\mathbb{P}$ while $k(x,y)=e^{i\langle x,y\rangle}$, $x,y\in\mathbb{R}^d$ yields the characteristic function of $\mathbb{P}$, which are known to characterize $\mathbb{P}$ completely (note that $k(x,y)=e^{i\langle x,y\rangle}$ is however not a valid reproducing kernel). Here $i:=\sqrt{-1}$. Therefore, mapping random variables, X into a suitable RKHS (as $\Phi(X)$) provides a powerful and straightforward method of dealing with the higher-order statistics of the variables.

Having provided an interpretation of the embedding in (1.2) through the framework of learning in RKHS and its advantage, in the following section, we summarize our contributions.

1.2 Summary of Contributions

From the above discussion, it should be clear that by appropriately choosing k (or equivalently Φ),

$$\int_{\mathcal{X}}k(\cdot,x)\,d\mathbb{P}(x)\in\mathcal{H}$$

completely characterizes $\mathbb{P}$. Therefore, a natural question to answer is, "for what k is the embedding in (1.2) injective?"—such kernels are defined as *characteristic kernels* [30]. The injectivity of (1.2) is critical in applications like two-sample tests where we need $\gamma_k(\mathbb{P},\mathbb{Q})=0$ if and only if $\mathbb{P}=\mathbb{Q}$. While various characterizations have been obtained in literature [30, 37, 80] for k to be characteristic—the drawback of these characterizations is that they are either difficult to check or may apply only on compact $\mathcal{X}$—in Chapter 3, we obtain easily checkable characterizations for characteristic kernels that are translation invariant (on $\mathbb{R}^d$ and $\mathbb{T}^d$) and radial on $\mathbb{R}^d$ (see Theorem 3.13, Corollary 3.16, Theorem 3.19 and Table 3.1). In particular, we show that a bounded continuous translation invariant kernel on $\mathbb{R}^d$ is characteristic if and only if the support of its Fourier transform is $\mathbb{R}^d$. Note that if k is characteristic then γ_k is a metric (and not just a pseudometric) on the space of probability measures. We show that $\gamma_k(\mathbb{P},\mathbb{Q})$ is the weighted L^2 distance between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$, with the weighting determined by the Fourier transform of the kernel (see Theorem 3.4). This chapter is based on the material published in [75, 78, 79].

In Chapter 4, we discuss how the characteristic property of a kernel is related to richness of the corresponding RKHS, where richness corresponds to how well the RKHS approximates certain target space of functions. In addition, we also discuss the relation of characteristic kernels to various notions of *positive definite (pd)* kernels like *strictly pd*, *integrally strictly pd* and *conditionally strictly pd* kernels (see Chapter 2 for their definitions). We refer the reader to Figure 4.1 for a summary of the relations between these various notions of pd kernels. This chapter is based on the material published in [78].

As mentioned above, the results in Chapters 3 and 4 provide conditions on k for which γ_k is a metric on the space of probability measures. Since many distance measures on probabilities have been studied in literature, with two popular families being integral probability metrics (IPMs) [55] and ϕ-divergences [1, 15], in Chapter 5, we discuss the advantages and disadvantages of γ_k over these families. In particular, we compare γ_k to these families on two respects: (a) ease of computation and estimation and (b) *strength* of the distance measure. We show that while $\gamma_k(\mathbb{P},\mathbb{Q})$ has a nice closed form expression (see (3.5)), it is not easily computable for all $\mathbb{P}$ and $\mathbb{Q}$ (which is also the case with IPMs and ϕ-divergences). Therefore, we approximate these distances based on finite samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, wherein we show that the estimator (or approximator) of γ_k is efficient to compute (see Theorems 5.1, 5.3 and 5.4), *strongly consistent* and exhibits fast convergence rate compared to those of IPMs and ϕ-divergences (see Corollary 5.12). Therefore, we argue that γ_k is well suited for statistical inference applications like two-sample tests than these other families of distances under consideration. Since there is an inherent connection between two-sample tests and *binary classification* problems, we show how γ_k is related to the *risk* associated with a particular binary classification problem (see Propositions 5.5 and 5.6). On a more theoretical front, we derive conditions on k under which γ_k metrizes the weak* (weak-star) topology on the space of probability measures—bounded continuous translation invariant characteristic kernels on $\mathbb{R}^d$ are shown to metrize the weak* topology—and thereby show that γ_k is *weaker* than popular distance measures like Kullbach-Leibler divergence, total variation distance, Wasserstein distance, etc. This chapter is based on the material published in [75, 76, 78].

Although we show in Chapter 5 that γ_k has many favorable properties for it to be used in applications like two-sample tests, one of its disadvantages as we show in Chapter 6 is that it is not clear how to choose an appropriate kernel for the problem at hand. To elaborate, suppose $k(x, y) = \exp(-\sigma\|x-y\|_2^2)$, $x, y \in \mathbb{R}^d$, $\sigma \in \mathbb{R}_{++}$, which is characteristic for any $\sigma \in \mathbb{R}_{++}$. Using k in γ_k, we have a family of distance measures on probabilities that is indexed by $\sigma \in \mathbb{R}_{++}$. Now, which of these distances should we consider as the distance between $\mathbb{P}$ and $\mathbb{Q}$ when γ_k is used in a two-sample test? We address this problem in Chapter 6 wherein we propose a new distance measure that generalizes to family of pd kernels. Depending on $\mathbb{P}$ and $\mathbb{Q}$, the proposed distance measure chooses an appropriate kernel, k^* so that $\gamma_{k^*}(\mathbb{P}, \mathbb{Q})$ can maximally differentiate between $\mathbb{P} \neq \mathbb{Q}$. We present experimental results of a two-sample test wherein it is shown that the proposed distance measure exhibits better performance in distinguishing between $\mathbb{P} \neq \mathbb{Q}$ than using γ_k with k being chosen heuristically (see Figure 6.1). This chapter is based on the material published in [75].

While Chapters 3–6 present various properties of the embedding in (1.2), i.e., RKHS embedding of probability measures, in Chapter 7, we extend this notion by embedding probability measures into a reproducing kernel Banach space (RKBS) [95]. Since RKBSs are generalization of RKHSs, we show that richer distances between probabilities can be obtained through this novel notion of embedding probabilities into an RKBS (see Example 7.16). We show that many of the results derived for the RKHS embeddings neatly extend to the RKBS embeddings (see Theorems 7.5, 7.6, 7.7 and 7.13). But, one drawback with RKBS embeddings is that the associated distance measure and its estimator do not exhibit a simple closed form unlike their RKHS counterparts. However, in some special cases (see Examples 7.16–7.18), we show that this drawback can be resolved. The material of this chapter is original and has not been published elsewhere.

Unless otherwise stated, throughout this dissertation we assume that $\mathcal{X}$ is a topological space.

2 Kernels and Reproducing Kernel Hilbert Spaces

In this chapter, we provide the necessary background on kernels and reproducing kernel Hilbert spaces, which will be required to understand and appreciate the results in the forthcoming chapters. The results in this chapter are collected from [81, Chapter 4] and [91]. For a comprehensive treatment on kernels and reproducing kernel Hilbert spaces, we refer the reader to [4, 66, 68, 81, 91].

The chapter is organized as follows. In Section 2.1, we define positive definite (pd) functions and kernels, discuss their properties and provide some examples. While we define the related notions of integrally pd and conditionally pd kernels in Section 2.2, the function space associated with pd kernels, called the reproducing kernel Hilbert space is introduced and discussed in Section 2.3.

2.1 Positive Definite Functions and Kernels

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *positive definite* (pd) if, for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, we have

$$\sum_{l,j=1}^{n} \alpha_l \alpha_j k(x_l, x_j) \geq 0. \tag{2.1}$$

Furthermore, k is said to be *strictly pd* if, for mutually distinct $x_1, \ldots, x_n \in \mathcal{X}$, equality in (2.1) only holds for $\alpha_1 = \cdots = \alpha_n = 0$. k is called *symmetric* if $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$. k is a called *kernel* on $\mathcal{X}$ if there exists a

Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$ we have

$$k(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}.$$

It can be shown that k is a kernel if and only if it is symmetric and pd [81, Theorem 4.16]. A simple example of kernel is the *dot-product kernel*, $k(x,y) = \langle x, y \rangle$, which is obtained by choosing $\mathcal{H} = \mathbb{R}^d$ and $\Phi(x) = x$.

Two popular classes of kernels that are considered in this dissertation are: (a) translation invariant kernels on $\mathbb{R}^d$ and $\mathbb{T}^d := [0, 2\pi)^d$ and (b) radial kernels on $\mathbb{R}^d$. A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to be *translation invariant* if $k(x,y) = \psi(x-y)$. The following theorem due to Bochner provides a characterization for ψ, which we quote from [91, Theorem 6.6].

Theorem 2.1 (Bochner). *A bounded continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if is the Fourier transform of a nonnegative finite Borel measure, Λ, i.e.,*

$$\psi(x) = \int_{\mathbb{R}^d} e^{-i\langle x, \omega \rangle} \, d\Lambda(\omega), \; x \in \mathbb{R}^d. \tag{2.2}$$

Therefore, a bounded continuous translation invariant function, k is a kernel on $\mathbb{R}^d$ if and only if

$$k(x,y) = \int_{\mathbb{R}^d} e^{-i\langle x-y, \omega \rangle} \, d\Lambda(\omega), \; x, y \in \mathbb{R}^d, \tag{2.3}$$

where $i := \sqrt{-1}$. Since k is real valued and symmetric, it is easy to see that Λ is a real-valued measure on $\mathbb{R}^d$ and $\Lambda(d\omega) = \Lambda(-d\omega)$. This means if $\widehat{\psi}$ is the Radon-Nikodym derivative of Λ w.r.t. the Lebesgue measure, i.e., $d\Lambda(\omega) = \widehat{\psi}(\omega)\, d\omega$, then $\widehat{\psi}(\omega) = \widehat{\psi}(-\omega) = \overline{\widehat{\psi}(\omega)}$, $\omega \in \mathbb{R}^d$, i.e., $\widehat{\psi}$ is real and even. Here $\overline{\widehat{\psi}(\omega)}$ represents the complex conjugate of $\widehat{\psi}(\omega)$.

A continuous translation invariant function, k is said to be a kernel on $\mathbb{T}^d := [0, 2\pi)^d$ if $k(x,y) = \psi((x-y)_{\mathrm{mod}\, 2\pi})$, where $\psi \in C(\mathbb{T}^d)$ is such that

$$\psi(x) = \sum_{n \in \mathbb{Z}^d} A_\psi(n) e^{i\langle x, n \rangle}, \; x \in \mathbb{T}^d, \tag{2.4}$$

with $A_\psi : \mathbb{Z}^d \to \mathbb{R}_+$, $A_\psi(-n) = A_\psi(n)$ and $\sum_{n \in \mathbb{Z}^d} A_\psi(n) < \infty$. Similar to Theorem 2.1, which is Bochner's theorem on $\mathbb{R}^d$, (2.4) can be seen as Bochner's theorem on $\mathbb{T}^d$.

$k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ is said to be *radial* on $\mathbb{R}^d$ if $k(x,y)=\eta(\|x-y\|_2^2)$. Similar to Theorem 2.1, the following theorem due to Schoenberg provides a characterization for the positive definiteness of k, which we quote from [91, Corollary 7.12 and Theorem 7.13].

Theorem 2.2 (Schoenberg)**.** *A bounded continuous function, $k(x,y)=\eta(\|x-y\|_2^2)$ is pd on $\mathbb{R}^d$ if and only if there exists a nonnegative finite Borel measure, ν on $[0,\infty)$ such that*

$$\eta(r)=\int_{[0,\infty)} e^{-rt}\,d\nu(t),\ x\in\mathbb{R}^d, \tag{2.5}$$

for all $r>0$.

The following proposition shows that a radial kernel is also translation invariant on $\mathbb{R}^d$.

Proposition 2.3. *If k is a radial pd kernel on $\mathbb{R}^d$, then it is also translation invariant.*

Proof. Let k be radial on $\mathbb{R}^d$. Then

$$k(x,y)=\psi(x-y):=\int_{[0,\infty)} e^{-t\|x-y\|_2^2}\,d\nu(t),\ x,y\in\mathbb{R}^d,$$

where ν is a finite nonnegative Borel measure on $[0,\infty)$. Since

$$e^{-t\|x-y\|_2^2}=\int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle}(4\pi t)^{-d/2}e^{-\|\omega\|_2^2/4t}\,d\omega,$$

we have $\psi(x)=\int_{\mathbb{R}^d} e^{-i\langle x,\omega\rangle}\phi(\omega)\,d\omega$, where $\phi(\omega)=\int_{[0,\infty)}(4\pi t)^{-d/2}e^{-\|\omega\|_2^2/4t}\,d\nu(t)$. It is easy to check that $\phi(\omega)\ge 0,\ \forall\,\omega\in\mathbb{R}^d$ and $\phi\in L^1(\mathbb{R}^d)$. Therefore k satisfies (2.3), which means k is translation invariant on $\mathbb{R}^d$. □

In the following, we provide popular examples of translation invariant kernels on $\mathbb{R}^d$ and $\mathbb{T}^d$.

Example 2.4 (Translation invariant kernels on $\mathbb{R}^d$)**.** *Let $d\Lambda(\omega)=(2\pi)^{-d/2}\widehat{\psi}(\omega)\,d\omega$. The following are translation invariant kernels on $\mathbb{R}^d$ as they satisfy (2.3) with $\widehat{\psi}\ge 0$ and $\widehat{\psi}\in L^1(\mathbb{R}^d)$. Here $\widehat{\psi}$ is the Fourier transform of ψ—see (C.2) and (C.3).*

(1) Gaussian kernel:

$$k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right),\ \sigma > 0,$$
$$\widehat{\psi}(\omega) = \sigma^d \exp\left(-\frac{\sigma^2\|\omega\|_2^2}{2}\right).$$

Note that the Gaussian kernel is also radial on $\mathbb{R}^d$ as it can be obtained by choosing $\nu = \delta_{\frac{1}{2\sigma^2}}$ in (2.5), where δ_x represents the Dirac measure at x.

(2) Laplacian kernel:

$$k(x,y) = \exp\left(-\sigma\|x-y\|_1\right),\ \sigma > 0,$$
$$\widehat{\psi}(\omega) = \left(\frac{2}{\pi}\right)^{d/2} \prod_{j=1}^{d} \frac{\sigma}{\sigma^2 + \omega_j^2},$$

where $\omega = (\omega_1, \ldots, \omega_d)$.

(3) B_{2n+1}-spline kernel [68]:

$$k(x,y) = \prod_{j=1}^{d} *_1^{2n+2} \mathbb{1}_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x_j - y_j),$$
$$\widehat{\psi}(\omega) = \prod_{j=1}^{d} \frac{4^{n+1}}{\sqrt{2\pi}} \frac{\sin^{2n+2}\left(\frac{\omega_j}{2}\right)}{\omega_j^{2n+2}},$$

where $*_1^{2n+2}$ represents the $(2n+2)$-fold convolution, $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d)$ and $\omega = (\omega_1, \ldots, \omega_d)$. Choosing $n = 0$ gives the B_1-spline kernel,

$$k(x,y) = \prod_{j=1}^{d} (1 - |x_j - y_j|) \mathbb{1}_{[-1,1]}(x_j - y_j),$$
$$\widehat{\psi}(\omega) = \prod_{j=1}^{d} \frac{4}{\sqrt{2\pi}} \frac{\sin^2(\omega_j/2)}{\omega_j^2}.$$

(4) Inverse multiquadratic kernel:

$$k(x,y) = (c^2 + \|x-y\|_2^2)^{-\beta},\ c > 0,\ \beta > \frac{d}{2},$$
$$\widehat{\psi}(\omega) = \frac{2^{1-\beta}}{\Gamma(\beta)} \left(\frac{\|\omega\|_2}{c}\right)^{\beta - d/2} K_{d/2-\beta}(c\|\omega\|_2),$$

where Γ is the Gamma function and K_ν is a modified Bessel function of the third kind of order $\nu \in \mathbb{R}$ [91, Theorem 6.13]. It is easy to check that the inverse multiquadratic kernel is also radial on $\mathbb{R}^d$ as it can be obtained by choosing $d\nu(t) = \frac{1}{\Gamma(\beta)} t^{\beta-1} e^{-c^2 t}\, dt$ in (2.5).

(5) *Matérn kernel [63, Section 4.2.1]:*

$$k(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|x-y\|_2}{\sigma}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\|x-y\|_2}{\sigma}\right),\ \nu > 0,\ \sigma > 0,$$

$$\widehat{\psi}(\omega) = \frac{2^{d+\nu}\pi^{d/2}\Gamma(\nu+d/2)\nu^{\nu}}{\Gamma(\nu)\sigma^{2\nu}} \left(\frac{2\nu}{\sigma^2} + 4\pi^2\|\omega\|_2^2\right)^{-(\nu+d/2)},$$

where ν controls the smoothness of k. The case of $\nu = \frac{1}{2}$ in the Matérn class gives the exponential kernel, $k(x,y) = \exp(-\|x-y\|_1/\sigma)$, while $\nu \to \infty$ gives the Gaussian kernel. Note that $\widehat{\psi}(x-y)$ is actually the inverse multiquadratic kernel.

(6) *Sinc kernel:*

$$k(x,y) = \prod_{j=1}^{d} \frac{\sin \sigma(x_j - y_j)}{x_j - y_j},\ \sigma > 0,$$

$$\widehat{\psi}(\omega) = \left(\frac{\pi}{2}\right)^{d/2} \prod_{j=1}^{d} \mathbb{1}_{[-\sigma,\sigma]}(\omega_j).$$

(7) *Sinc-squared kernel:*

$$k(x,y) = \prod_{j=1}^{d} \frac{\sin^2 \frac{x_j - y_j}{2}}{(x_j - y_j)^2},$$

$$\widehat{\psi}(\omega) = \frac{(2\pi)^{d/2}}{4^d} \prod_{j=1}^{d} (1 - |\omega_j|)\mathbb{1}_{[-1,1]}(\omega_j).$$

Example 2.5 (Translation invariant kernels on $\mathbb{T}$). *The following are translation invariant kernels on $\mathbb{T}$ as they satisfy (2.4) with $A_\psi(n) \geq 0, \forall\, n \in \mathbb{Z}$, $A_\psi(n) = A_\psi(-n)$, $n \in \mathbb{Z}$ and $\sum_{n\in\mathbb{Z}} A_\psi(n) < \infty$.*

(1) *Poisson kernel [11, 80, 87]:*

$$k(x,y) = \frac{1-\sigma^2}{\sigma^2 - 2\sigma\cos(x-y) + 1},\ 0 < \sigma < 1,$$

$$A_\psi(n) = \sigma^{|n|}.$$

(2) Dirichlet kernel [11, 68]:

$$\begin{aligned} k(x,y) &= \frac{\sin\frac{(2l+1)(x-y)}{2}}{\sin\frac{(x-y)}{2}},\ l\in\mathbb{N},\\ A_\psi(n) &= \mathbb{1}_D(n), \end{aligned}$$

where $D := \{0, \pm 1, \dots, \pm l\}$.

(3) Fejér kernel [11]:

$$\begin{aligned} k(x,y) &= \frac{1}{l+1}\frac{\sin^2\frac{(l+1)(x-y)}{2}}{\sin^2\frac{(x-y)}{2}},\ l\in\mathbb{N},\\ A_\psi(n) &= \left(1-\frac{|n|}{l+1}\right)\mathbb{1}_D(n). \end{aligned}$$

Theorem 6.11 in [91] shows that a continuous function $\psi \in L^1(\mathbb{R}^d)$ is strictly pd if and only if ψ is bounded and $\widehat{\psi}$ is nonnegative and nonvanishing. This means the kernels in Example 2.4 are strictly pd (however, the strict positive definiteness of the sinc kernel does not follow from this result as the sinc kernel is not integrable).

2.2 Integrally and Conditionally Positive Definite Functions

Apart from pd kernels, two related notions that appear in this dissertation (particularly in Chapters 3 and 4) are: (a) integrally strictly pd and (b) conditionally strictly pd functions.

A measurable, symmetric and bounded function, k is said to be *integrally strictly pd* if

$$\iint_{\mathcal{X}} k(x,y)\,d\mu(x)\,d\mu(y) > 0,\ \forall\,\mu\in M_b(\mathcal{X})\backslash\{0\}, \tag{2.6}$$

where $M_b(\mathcal{X})$ is the set of all finite signed Borel measures on $\mathcal{X}$. This definition is a generalization of *integrally strictly positive definite functions* on $\mathbb{R}^d$ [82, Section 6]: $\iint_{\mathbb{R}^d} k(x,y)f(x)f(y)\,dx\,dy > 0$ for all $f \in L^2(\mathbb{R}^d)$, which is the strict positive definiteness of the integral operator given by the kernel. The following result shows that integrally strictly pd functions are strictly pd kernels, while the converse is not true which follows from [81, Proposition 4.60, Theorem 4.62].

Proposition 2.6. *If k is integrally strictly pd, then it is strictly pd.*

Proof. Suppose k is not strictly pd. This means for some $n \in \mathbb{N}$ and for mutually distinct $x_1, \ldots, x_n \in \mathcal{X}$, there exists $\mathbb{R} \ni \alpha_j \neq 0$ for some $j \in \{1, \ldots, n\}$ such that $\sum_{j,l=1}^{n} \alpha_j \alpha_l k(x_j, x_l) = 0$. By defining $\mu = \sum_{j=1}^{n} \alpha_j \delta_{x_j}$, it is easy to see that there exists $\mu \neq 0$ such that $\iint_{\mathcal{X}} k(x,y)\, d\mu(x)\, d\mu(y) = 0$, which means k is not integrally strictly pd. Therefore, if k is integrally strictly pd, then it is strictly pd. Here, δ_x represents the Dirac measure at $x \in \mathcal{X}$. □

Examples of integrally strictly pd kernels on $\mathbb{R}^d$ include the Gaussian, Laplacian, B_{2n+1}-splines, inverse multiquadratic, Matérn kernels, etc., which follows from the following result.

Proposition 2.7. *Suppose $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$ where ψ is a bounded continuous pd function on $\mathbb{R}^d$. Then k is integrally strictly pd if* $\operatorname{supp}(\Lambda) = \mathbb{R}^d$*, where Λ is defined in (2.2).*[1]

Proof. Consider $\iint_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y)$ for any $\mu \in M_b(\mathbb{R}^d)$ with $k(x,y) = \psi(x-y)$, where $M_b(\mathbb{R}^d)$ is the set of all finite signed Borel measures on $\mathbb{R}^d$.

$$\begin{aligned}
B &:= \iint_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y) \\
&= \iint_{\mathbb{R}^d} \psi(x-y)\, d\mu(x)\, d\mu(y) \\
&\overset{(d)}{=} \iiint_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle}\, d\Lambda(\omega)\, d\mu(x)\, d\mu(y) \\
&\overset{(e)}{=} \iint_{\mathbb{R}^d} e^{-i\langle x,\omega\rangle}\, d\mu(x) \int_{\mathbb{R}^d} e^{i\langle y,\omega\rangle}\, d\mu(y)\, d\Lambda(\omega) \\
&\overset{(f)}{=} \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\widehat{\mu}(\omega)}\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \left|\widehat{\mu}(\omega)\right|^2\, d\Lambda(\omega), \qquad (2.7)
\end{aligned}$$

where Bochner's theorem (Theorem 2.1) is invoked in (d), Fubini's theorem (Theorem C.1) in (e) and (C.5) in (f). If $\operatorname{supp}(\Lambda) = \mathbb{R}^d$, then it is clear that $B > 0$. □

[1] See (C.1) for the definition of support of a Borel measure, Λ.

Note that the Gaussian, Laplacian, B_{2n+1}-splines, inverse multiquadratic, Matérn kernels, etc., satisfy $\text{supp}(\Lambda) = \text{supp}(\widehat{\psi}) = \mathbb{R}^d$ and are therefore integrally strictly pd. However, it can be shown that the sinc and sinc-squared kernels are not integrally strictly pd (see Theorem 3.13). The following result provides a characterization for integrally strictly pd kernels when k is translation invariant on $\mathcal{X} = \mathbb{T}^d$.

Proposition 2.8. *Suppose $k(x,y) = \psi((x-y)_{mod\,2\pi})$, $x,y \in \mathbb{T}^d$ where ψ is a continuous pd function on $\mathbb{T}^d$. Then k is integrally strictly pd if and only if $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}^d$, where A_ψ is defined in (2.4).*

Proof. See Proposition 4.5. □

Therefore, the Poisson kernel on $\mathbb{T}$ is integrally strictly pd, while the Dirichlet and Fejér kernels are not.

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *conditionally pd* if, for all $n \geq 2$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ with $\sum_{j=1}^n \alpha_j = 0$ and all $x_1, \ldots, x_n \in X$, we have

$$\sum_{l,j=1}^{n} \alpha_l \alpha_j k(x_l, x_j) \geq 0.$$

Furthermore, k is said to be *conditionally strictly pd* if, for mutually distinct $x_1, \ldots, x_n \in \mathcal{X}$, equality in (2.1) only holds for $\alpha_1 = \cdots = \alpha_n = 0$. From the definitions of strictly pd and conditionally strictly pd functions, it is clear that strictly pd kernels are conditionally strictly pd but not vice-versa.

2.3 The Reproducing Kernel Hilbert Space of a Kernel

In this section, we introduce reproducing kernel Hilbert spaces (RKHSs) and describe their relation to kernels.

Definition 2.9. *Let $\mathcal{X} \neq \emptyset$ and $\mathcal{H}$ be an Hilbert space of functions over $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if $k(\cdot, x) \in \mathcal{H}$ for all*

$x \in \mathcal{X}$ and the reproducing property

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$$

holds for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$. The space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) over $\mathcal{X}$ if for all $x \in \mathcal{X}$ the Dirac functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by

$$\delta_x(f) := f(x),\ f \in \mathcal{H},$$

is continuous.

Since $L^2(\mathbb{R}^d)$ does not consist of functions, it is not an RKHS. Lemma 4.19 in [81] shows that reproducing kernels are kernels, where $\Phi(x) = k(\cdot, x)$, $x \in \mathcal{X}$, i.e., $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}$, $x, y, \in \mathcal{X}$. The Moore-Aronszajn theorem states that there is a one-to-one relation between kernels and RKHSs, i.e., every RKHS has a unique reproducing kernel and every kernel has a unique RKHS (see Theorems 4.20 and 4.21 in [81]).

A nice interpretation for RKHS can be obtained through the following characterization, quoted from [91, Theorem 10.12], if k is translation invariant on $\mathbb{R}^d$.

Theorem 2.10 ([91]). *Suppose $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$, where $\psi \in C(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ is a real-valued strictly pd function. Define*

$$\mathcal{H} := \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \ : \ \frac{\widehat{f}}{\sqrt{\widehat{\psi}}} \in L^2(\mathbb{R}^d) \right\} \tag{2.8}$$

and equip this space with the inner product

$$\langle f, g \rangle_{\mathcal{H}} := \frac{1}{(2\pi)^{d/2}} \left\langle \frac{\widehat{f}}{\sqrt{\widehat{\psi}}}, \frac{\widehat{g}}{\sqrt{\widehat{\psi}}} \right\rangle_{L^2(\mathbb{R}^d)} = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{\widehat{f}(\omega)\overline{\widehat{g}(\omega)}}{\widehat{\psi}(\omega)}\, d\omega. \tag{2.9}$$

Then $\mathcal{H}$ is a real Hilbert space with k as the reproducing kernel.

Suppose k is a Matérn kernel on $\mathbb{R}^d$, which by Theorem 2.10 means that

$$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \ : \ \widehat{f}(\cdot) \left(1 + \| \cdot \|_2^2\right)^{s/2} \in L^2(\mathbb{R}^d) \right\} \tag{2.10}$$

is an RKHS with the Matérn kernel as its r.k. Since (2.10) is a Sobolev space of order s for $s > d/2$, the RKHS in (2.8) can be seen as a generalization of Sobolev spaces on $\mathbb{R}^d$.

3 Characteristic Kernels and Maximum Mean Discrepancy

In Chapter 1, we have motivated and introduced the notion of embedding a Borel probability measure, $\mathbb{P}$—defined on a topological space, $\mathcal{X}$—into an RKHS, $(\mathcal{H}, k)$ as

$$\mathbb{P} \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x), \tag{3.1}$$

using which the *maximum mean discrepancy* (MMD) is defined as the RKHS distance between the embeddings of probability measures $\mathbb{P}$ and $\mathbb{Q}$, given by

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \left\| \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}}.$$

For γ_k to be useful in practice, e.g., in applications like two-sample tests, it is requisite that the embedding in (3.1) is injective so that γ_k is a metric on $M_+^1(\mathcal{X})$, the set of Borel probability measures on $\mathcal{X}$—note that irrespective of whether (3.1) is injective or not, γ_k is a pseudometric on $M_+^1(\mathcal{X})$.[2] The main focus of this chapter is to study the conditions on k for which (3.1) is injective.

The chapter is organized as follows. First, in order to obtain a better understanding of γ_k, in Section 3.1, we present results which provide different interpretations of γ_k. Next, in Section 3.2, we present our main results on the characterization of k for which (3.1) is injective—such kernels are defined to be *characteristic kernels*. In Section 3.3, we present a property of γ_k wherein we show

[2] Given a set $\mathcal{X}$, a *metric* for $\mathcal{X}$ is a function $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that *(i)* $\forall\, x$, $\rho(x, x) = 0$, *(ii)* $\forall\, x, y$, $\rho(x, y) = \rho(y, x)$, *(iii)* $\forall\, x, y, z$, $\rho(x, z) \le \rho(x, y) + \rho(y, z)$, and *(iv)* $\rho(x, y) = 0 \Rightarrow x = y$. A semi-metric only satisfies *(i)*, *(ii)* and *(iv)*. A pseudometric only satisfies *(i)*-*(iii)* of the properties of a metric. Unlike a metric space $(\mathcal{X}, \rho)$, points in a pseudometric space need not be distinguishable: one may have $\rho(x, y) = 0$ for $x \neq y$.

that even if k is characteristic, there exist two distinct probability measures, $\mathbb{P}$ and $\mathbb{Q}$ for any $\varepsilon > 0$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) < \varepsilon$. In other words, though a characteristic kernel distinguishes between distinct $\mathbb{P}$ and $\mathbb{Q}$, there can exist distributions that are difficult to distinguish. This property of γ_k is significant in applications like two-sample tests where $\mathbb{P}$ and $\mathbb{Q}$ are estimated from finite samples and the distance between them (i.e., the estimates of $\mathbb{P}$ and $\mathbb{Q}$) may not be statistically significant.

3.1 Interpretation of MMD

While one can start with the definition of the embedding in (3.1) and then study its associated metric, γ_k, we show in Proposition 3.2 that such an embedding can be obtained by relating $\gamma_k(\mathbb{P},\mathbb{Q})$ to the *integral probability metric* (IPM) between $\mathbb{P}$ and $\mathbb{Q}$, defined as

$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{X}} f\, d\mathbb{P} - \int_{\mathcal{X}} f\, d\mathbb{Q} \right|, \tag{3.2}$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on $\mathcal{X}$. In particular we show that $\gamma_{\mathcal{F}_k}(\mathbb{P},\mathbb{Q}) = \gamma_k(\mathbb{P},\mathbb{Q})$, where $\mathcal{F}_k := \{f : \|f\|_{\mathcal{H}} \le 1\}$. See Chapter 5 for a detailed discussion on the advantages of γ_k over other IPMs. We would like to mention that a result similar to Proposition 3.2 was also derived by [37] and [72], but here we prove it rigorously. To prove Proposition 3.2, we need the following supplementary result.

Lemma 3.1. *Let k be a measurable and bounded pd kernel on a measurable space $\mathcal{X}$ and let $\mathcal{H}$ be its associated RKHS. Suppose μ be a finite signed measure on $\mathcal{X}$ such that $\int_{\mathcal{X}} \sqrt{k(x,x)}\, d|\mu|(x) < \infty$. Then, for any $f \in \mathcal{H}$, we have*

$$\int_{\mathcal{X}} f(x)\, d\mu(x) = \int_{\mathcal{X}} \langle f, k(\cdot,x)\rangle_{\mathcal{H}}\, d\mu(x) = \left\langle f, \int_{\mathcal{X}} k(\cdot,x)\, d\mu(x) \right\rangle_{\mathcal{H}}. \tag{3.3}$$

Proof. Let $T_\mu : \mathcal{H} \to \mathbb{R}$ be a linear functional defined as $T_\mu[f] := \int_{\mathcal{X}} f(x)\, d\mu(x)$, with

$$\|T_\mu\| := \sup\left\{ \frac{|T_\mu[f]|}{\|f\|_{\mathcal{H}}} : 0 \ne f \in \mathcal{H} \right\}.$$

Consider

$$|T_\mu[f]| = \left| \int_{\mathcal{X}} f(x)\, d\mu(x) \right| \le \int_{\mathcal{X}} |f(x)|\, d|\mu|(x) = \int_{\mathcal{X}} |\langle f, k(\cdot,x)\rangle_{\mathcal{H}}|\, d|\mu|(x)$$

$$\overset{(\star)}{\leq} \int_{\mathcal{X}} \sqrt{k(x,x)}\|f\|_{\mathcal{H}}\, d|\mu|(x),$$

which implies $\|T_\mu\| < \infty$, that is, T_μ is a bounded linear functional on $\mathcal{H}$—in $(\star)$, we used the fact that $\langle f, k(\cdot,x)\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}\|k(\cdot,x)\|_{\mathcal{H}} = \|f\|_{\mathcal{H}}\sqrt{\langle k(\cdot,x), k(\cdot,x)\rangle_{\mathcal{H}}} = \|f\|_{\mathcal{H}}\sqrt{k(x,x)}$. Therefore, by the Riesz representation theorem (Theorem C.2), there exists a unique $\lambda_\mu \in \mathcal{H}$ such that $T_\mu[f] = \langle f, \lambda_\mu\rangle_{\mathcal{H}}$, $\forall\, f \in \mathcal{H}$. Let $f = k(\cdot,u)$ for some $u \in \mathcal{X}$. Then, $T_\mu[k(\cdot,u)] = \langle k(\cdot,u), \lambda_\mu\rangle_{\mathcal{H}} = \lambda_\mu(u)$, which implies $\lambda_\mu = \int_{\mathcal{X}} k(\cdot,x)\, d\mu(x)$ and the result follows. □

Proposition 3.2. *Let* $\mathscr{P}_k(\mathcal{X}) := \left\{\mathbb{P} \in M^1_+(\mathcal{X}) : \int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mathbb{P}(x) < \infty\right\}$*, where* k *is measurable on* $\mathcal{X}$*. Then for any* $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_k(\mathcal{X})$*,*

$$\gamma_{\mathcal{F}_k}(\mathbb{P},\mathbb{Q}) = \gamma_k(\mathbb{P},\mathbb{Q}).$$

Proof. By Lemma 3.1, we have $\int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) = \langle f, \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{P}(x)\rangle_{\mathcal{H}}$ for any $\mathbb{P} \in \mathscr{P}_k(\mathcal{X})$. Therefore,

$$\begin{aligned}\gamma_{\mathcal{F}_k}(\mathbb{P},\mathbb{Q}) &= \sup_{\|f\|_{\mathcal{H}}\leq 1} \left|\int_{\mathcal{X}} f(x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} f(x)\, d\mathbb{Q}(x)\right| \\ &= \sup_{\|f\|_{\mathcal{H}}\leq 1} \left|\left\langle f, \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{Q}(x)\right\rangle_{\mathcal{H}}\right| \\ &= \left\|\int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{Q}(x)\right\|_{\mathcal{H}} = \gamma_k(\mathbb{P},\mathbb{Q}).\end{aligned}$$

Note that this holds for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_k(\mathcal{X})$. □

Proposition 3.2 shows that starting from an IPM in (3.2) and appropriately choosing $\mathcal{F}$ (in fact choosing $\mathcal{F} = \mathcal{F}_k$), one obtains γ_k and the embedding in (3.1), which hold for all $\mathbb{P} \in \mathscr{P}_k(\mathcal{X})$. However, in practice, especially in statistical inference applications, it is not possible to check whether $\mathbb{P} \in \mathscr{P}_k(\mathcal{X})$ as $\mathbb{P}$ is not known. Therefore, one would prefer to have a kernel such that

$$\int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mathbb{P}(x) < \infty, \forall\, \mathbb{P} \in M^1_+(\mathcal{X}). \tag{3.4}$$

The following proposition shows that (3.4) is equivalent to the kernel being bounded. Therefore, combining Propositions 3.2 and 3.3 shows that if k is measurable and bounded, then $\gamma_k(\mathbb{P},\mathbb{Q}) = \|\int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,x)\, d\mathbb{Q}(x)\|_{\mathcal{H}}$ for any $\mathbb{P}, \mathbb{Q} \in M^1_+(\mathcal{X})$.

Proposition 3.3. *Let f be a measurable function on $\mathcal{X}$. Then $\int_{\mathcal{X}} f(x)\,d\mathbb{P}(x) < \infty$ for all $\mathbb{P} \in M_+^1(\mathcal{X})$ if and only if f is bounded.*

Proof. One direction is straightforward because if f is bounded, then

$$\int_{\mathcal{X}} f(x)\,d\mathbb{P}(x) < \infty \text{ for all } \mathbb{P} \in M_+^1(\mathcal{X}).$$

Let us consider the other direction. Suppose f is not bounded. Then there exists a sequence $\{x_n\} \subset \mathcal{X}$ such that $f(x_n) \overset{n\to\infty}{\longrightarrow} \infty$. By taking a subsequence, if necessary, we can assume $f(x_n) > n^2$ for all n. Then, $A := \sum_{n=1}^{\infty} \frac{1}{f(x_n)} < \infty$. Define a probability measure $\mathbb{P}$ on $\mathcal{X}$ by $\mathbb{P} = \sum_{n=1}^{\infty} \frac{1}{Af(x_n)}\,\delta_{x_n}$, where δ_{x_n} is a Dirac measure at x_n. Then, $\int_{\mathcal{X}} f(x)\,d\mathbb{P}(x) = \frac{1}{A}\sum_{n=1}^{\infty} \frac{f(x_n)}{f(x_n)} = \infty$, which means if f is not bounded, then there exists a $\mathbb{P} \in M_+^1(\mathcal{X})$ such that $\int_{\mathcal{X}} f(x)\,d\mathbb{P}(x) = \infty$. □

Before presenting other interpretations of γ_k, in the following, we present a number of equivalent representations of γ_k, which will be helpful in its computation. [37] has shown that the reproducing property of k leads to

$$\begin{aligned}
\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \left\| \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{Q}(x) \right\|_{\mathcal{H}}^2 \\
&= \left\langle \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x), \int_{\mathcal{X}} k(\cdot,y)\,d\mathbb{P}(y) \right\rangle_{\mathcal{H}} \\
&\qquad + \left\langle \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{Q}(x), \int_{\mathcal{X}} k(\cdot,y)\,d\mathbb{Q}(y) \right\rangle_{\mathcal{H}} \\
&\qquad\qquad -2\left\langle \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x), \int_{\mathcal{X}} k(\cdot,y)\,d\mathbb{Q}(y) \right\rangle_{\mathcal{H}} \\
&\overset{(a)}{=} \iint_{\mathcal{X}} k(x,y)\,d\mathbb{P}(x)\,d\mathbb{P}(y) + \iint_{\mathcal{X}} k(x,y)\,d\mathbb{Q}(x)\,d\mathbb{Q}(y) \\
&\qquad\qquad -2\iint_{\mathcal{X}} k(x,y)\,d\mathbb{P}(x)\,d\mathbb{Q}(y) \qquad (3.5) \\
&= \iint_{\mathcal{X}} k(x,y)\,d(\mathbb{P}-\mathbb{Q})(x)\,d(\mathbb{P}-\mathbb{Q})(y), \qquad (3.6)
\end{aligned}$$

where (a) follows from (3.3). This means γ_k^2 is a straightforward sum of expectations of k, and can be computed easily, for example, using (3.5) either in closed form or using numerical integration techniques, depending on the choice of k, $\mathbb{P}$ and $\mathbb{Q}$. It is easy to show that, if k is a Gaussian kernel with $\mathbb{P}$ and $\mathbb{Q}$ being normal distributions on $\mathbb{R}^d$, then γ_k can be computed in a closed form (see Section 5.2.4 for examples).

In the following theorem, we prove three results which provide a nice interpretation for γ_k when $\mathcal{X} = \mathbb{R}^d$ and k is translation invariant, that is, $k(x,y) = \psi(x-y)$, where ψ is a pd function. We provide a detailed explanation to Theorem 3.4 in Remark 3.5.

Theorem 3.4 (Different interpretations of γ_k)**.** *(i) Let $\mathcal{X} = \mathbb{R}^d$ and $k(x,y) = \psi(x-y)$, where $\psi : \mathcal{X} \to \mathbb{R}$ is a bounded, continuous pd function. Then for any $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$,*

$$\gamma_k(\mathbb{P},\mathbb{Q}) = \sqrt{\int_{\mathbb{R}^d} |\phi_{\mathbb{P}}(\omega) - \phi_{\mathbb{Q}}(\omega)|^2 \, d\Lambda(\omega)} =: \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d,\Lambda)}, \tag{3.7}$$

where $\Lambda \in M_b^+(\mathbb{R}^d)$ is defined in (2.2). $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ represent the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$ respectively, and $M_b^+(\mathbb{R}^d)$ is the set of all nonnegative finite Borel measures on $\mathbb{R}^d$.

(ii) Suppose $\theta \in L^1(\mathbb{R}^d)$ is a continuous bounded pd function and $\int_{\mathbb{R}^d} \theta(x)\,dx = 1$. Let $\psi(x) := \psi_t(x) = t^{-d}\theta(t^{-1}x)$, $t > 0$. Assume that p and q are bounded uniformly continuous Radon-Nikodym derivatives of $\mathbb{P}$ and $\mathbb{Q}$ w.r.t. the Lebesgue measure, that is, $d\mathbb{P} = p\,dx$ and $d\mathbb{Q} = q\,dx$. Then,

$$\lim_{t\to 0} \gamma_k(\mathbb{P},\mathbb{Q}) = \|p-q\|_{L^2(\mathbb{R}^d)}. \tag{3.8}$$

In particular, if $|\theta(x)| \le C(1+\|x\|_2)^{-d-\varepsilon}$ for some $C, \varepsilon > 0$, then (3.8) holds for all bounded p and q (not necessarily uniformly continuous).

(iii) Suppose $\psi \in L^1(\mathbb{R}^d)$ and $\sqrt{\widehat{\psi}} \in L^1(\mathbb{R}^d)$. Then,

$$\gamma_k(\mathbb{P},\mathbb{Q}) = (2\pi)^{-d/4}\|\Phi * \mathbb{P} - \Phi * \mathbb{Q}\|_{L^2(\mathbb{R}^d)}, \tag{3.9}$$

where $\Phi := \left(\sqrt{\widehat{\psi}}\right)^{\vee}$ and $d\Lambda(\omega) = (2\pi)^{-d/2}\widehat{\psi}(\omega)\,d\omega$. Here, $\Phi * \mathbb{P}$ represents the convolution of Φ and $\mathbb{P}$.

Proof. (i) Let us consider (2.7) with $\mu := \mathbb{P} - \mathbb{Q}$ which yields (3.6). Since $\widehat{\mathbb{P}} = \overline{\phi_{\mathbb{P}}}$ (see (C.5)), we have $\widehat{\mu} = \overline{\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}}$ and the result follows from (2.7).

(ii) Consider (3.5) with $k(x,y) = \psi_t(x-y)$,

$$\begin{aligned}\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \iint_{\mathbb{R}^d} \psi_t(x-y)p(x)p(y)\,dx\,dy + \iint_{\mathbb{R}^d} \psi_t(x-y)q(x)q(y)\,dx\,dy \\ &\qquad -2\iint_{\mathbb{R}^d} \psi_t(x-y)p(x)q(y)\,dx\,dy \\ &= \int_{\mathbb{R}^d} (\psi_t * p)(x)p(x)\,dx + \int_{\mathbb{R}^d} (\psi_t * q)(x)q(x)\,dx \\ &\qquad -2\int_{\mathbb{R}^d} (\psi_t * q)(x)p(x)\,dx, \end{aligned} \tag{3.10}$$

where $(\psi_t * p)(x) := \int_{\mathbb{R}^d} \psi_t(x-y)p(y)\,dy$ is the convolution of ψ_t and p. Note that $\lim_{t\to 0}\int_{\mathbb{R}^d}(\psi_t * p)(x)p(x)\,dx = \int_{\mathbb{R}^d}\lim_{t\to 0}(\psi_t * p)(x)p(x)\,dx$, by invoking the dominated convergence theorem [26, Theorem 3.24]. Since p is bounded and uniformly continuous, by [26, Theorem 8.14], we have $p * \psi_t \to p$ uniformly as $t \to 0$, which means $\lim_{t\to 0}\int_{\mathbb{R}^d}(\psi_t * p)(x)p(x)\,dx = \int_{\mathbb{R}^d} p^2(x)\,dx$. Using this in (3.10), we have

$$\lim_{t\to 0}\gamma_k^2(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d}(p^2(x)+q^2(x)-2p(x)q(x))\,dx = \|p-q\|^2_{L^2(\mathbb{R}^d)}.$$

Suppose $|\theta(x)| \le (1+\|x\|_2)^{-d-\varepsilon}$ for some $C, \varepsilon > 0$. Since $p \in L^1(\mathbb{R}^d)$, by [26, Theorem 8.15], we have $(p*\psi_t)(x) \to p(x)$ as $t \to 0$ for almost every x. Therefore $\lim_{t\to 0}\int_{\mathbb{R}^d}(\psi_t * p)(x)p(x)\,dx = \int_{\mathbb{R}^d} p^2(x)\,dx$ and the result follows.

(iii) Since ψ is pd, $\widehat{\psi}$ is nonnegative and therefore $\sqrt{\widehat{\psi}}$ is valid. Since $\sqrt{\widehat{\psi}} \in L^1(\mathbb{R}^d)$, Φ exists. Define $\phi_{\mathbb{P},\mathbb{Q}} := \phi_{\mathbb{P}} - \phi_{\mathbb{Q}}$ and $\Phi * \mathbb{P} := \int_{\mathbb{R}^d}\Phi(\cdot - y)\,d\mathbb{P}(y)$. Now, consider

$$\begin{aligned}\|\Phi * (\mathbb{P}-\mathbb{Q})\|^2_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \left|(\Phi * (\mathbb{P}-\mathbb{Q}))(x)\right|^2\,dx \\ &= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\Phi(x-y)\,d(\mathbb{P}-\mathbb{Q})(y)\right|^2\,dx \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\iint_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)}\,e^{i\langle x-y,\omega\rangle}\,d\omega\;d(\mathbb{P}-\mathbb{Q})(y)\right|^2\,dx \\ &\overset{(c)}{=} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)}(\overline{\phi_{\mathbb{P}}(\omega)}-\overline{\phi_{\mathbb{Q}}(\omega)})\,e^{i\langle x,\omega\rangle}\,d\omega\right|^2\,dx \\ &= \frac{1}{(2\pi)^d}\iiint_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)\widehat{\psi}(\xi)}\,\overline{\phi_{\mathbb{P},\mathbb{Q}}(\omega)}\,\phi_{\mathbb{P},\mathbb{Q}}(\xi)\,e^{i\langle\omega-\xi,x\rangle}\,d\omega\,d\xi\,dx \\ &\overset{(d)}{=} \iint_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)\widehat{\psi}(\xi)}\,\overline{\phi_{\mathbb{P},\mathbb{Q}}(\omega)}\,\phi_{\mathbb{P},\mathbb{Q}}(\xi)\left[\int_{\mathbb{R}^d}\frac{e^{i\langle\omega-\xi,x\rangle}}{(2\pi)^d}\,dx\right]\,d\omega\,d\xi \\ &= \iint_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)}\sqrt{\widehat{\psi}(\xi)}\,\overline{\phi_{\mathbb{P},\mathbb{Q}}(\omega)}\,\phi_{\mathbb{P},\mathbb{Q}}(\xi)\,\delta(\omega-\xi)\,d\omega\,d\xi\end{aligned}$$

$$= \int_{\mathbb{R}^d} \widehat{\psi}(\omega) \left|\phi_{\mathbb{P}}(\omega) - \phi_{\mathbb{Q}}(\omega)\right|^2 \, d\omega = (2\pi)^{d/2}\gamma_k^2(\mathbb{P},\mathbb{Q}),$$

where (c) and (d) are obtained by invoking Fubini's theorem (Theorem C.1). □

Remark 3.5. *(a) (3.7) shows that γ_k is the L^2-distance between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$ computed w.r.t. the nonnegative finite Borel measure, Λ, which is the Fourier transform of ψ. If $\psi \in L^1(\mathbb{R}^d)$, then (3.7) rephrases the well known fact (see (2.9)) that for any $f \in \mathcal{H}$,*

$$\|f\|^2_{\mathcal{H}} = \int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega)|^2}{\widehat{\psi}(\omega)} \, d\omega. \tag{3.11}$$

Choosing $f = (\mathbb{P}-\mathbb{Q}) * \psi$ in (3.11) yields $\widehat{f} = (\phi_{\mathbb{P}} - \phi_{\mathbb{Q}})\widehat{\psi}$ and therefore the result in (3.7).

(b) Suppose $d\Lambda(\omega) = (2\pi)^{-d}\,d\omega$. Assume $\mathbb{P}$ and $\mathbb{Q}$ have p and q as Radon-Nikodym derivatives w.r.t. the Lebesgue measure, that is, $d\mathbb{P} = p\,dx$ and $d\mathbb{Q} = q\,dx$. Using these in (3.7), it can be shown that $\gamma_k(\mathbb{P},\mathbb{Q}) = \|p-q\|_{L^2(\mathbb{R}^d)}$. However, this result should be interpreted in a limiting sense as mentioned in Theorem 3.4(ii) because the choice of $d\Lambda(\omega) = (2\pi)^{-d}\,d\omega$ implies $\psi(x) = \delta(x)$, which does not satisfy the conditions of Theorem 3.4(i). It can be shown that $\psi(x) = \delta(x)$ is obtained in a limiting sense [26, Proposition 9.1]: $\psi_t \to \delta$ in $\mathcal{D}'_d$ as $t \to 0$.

(c) Choosing $\theta(x) = (2\pi)^{-d/2}e^{-\|x\|_2^2/2}$ in Theorem 3.4(ii) corresponds to ψ_t being a Gaussian kernel (with appropriate normalization such that $\int_{\mathbb{R}^d}\psi_t(x)\,dx = 1$). Therefore, (3.8) shows that as the bandwidth, t of the Gaussian kernel approaches zero, γ_k approaches the L^2-distance between the densities p and q. The same result also holds for choosing ψ_t as the Laplacian kernel, B_{2n+1}-spline, inverse multiquadratic, etc. Therefore, $\gamma_k(\mathbb{P},\mathbb{Q})$ can be seen as a generalization of the L^2-distance between probability measures, $\mathbb{P}$ and $\mathbb{Q}$.

(d) The result in (3.8) holds if p and q are bounded and uniformly continuous. Since any condition on $\mathbb{P}$ and $\mathbb{Q}$ is usually difficult to check in statistical applications, it is better to impose conditions on ψ rather than on $\mathbb{P}$ and $\mathbb{Q}$. In Theorem 3.4(ii), by imposing additional conditions on ψ_t, the result in (3.8) is shown to hold for all $\mathbb{P}$ and $\mathbb{Q}$ with bounded densities p and q. The condition, $|\theta(x)| \le C(1+\|x\|_2)^{-d-\varepsilon}$

for some $C, \varepsilon > 0$, is, for example, satisfied by the inverse multiquadratic kernel, $\theta(x) = \widetilde{C}(1+\|x\|_2^2)^{-\tau}$, $x \in \mathbb{R}^d$, $\tau > d/2$, where $\widetilde{C} = \left(\int_{\mathbb{R}^d}(1+\|x\|_2^2)^{-\tau}\,dx\right)^{-1}$.

(e) (3.9) shows that γ_k is proportional to the L^2-distance between $\Phi * \mathbb{P}$ and $\Phi * \mathbb{Q}$. Let Φ be such that Φ is nonnegative and $\Phi \in L^1(\mathbb{R}^d)$. Then, defining $\widetilde{\Phi} := \left(\int_{\mathbb{R}^d}\Phi(x)\,dx\right)^{-1}\Phi = \Phi/\sqrt{\widehat{\psi}(0)} = \left(\int_{\mathbb{R}^d}\psi(x)\,dx\right)^{-1/2}\Phi$ and using this in (3.9), we have

$$\gamma_k(\mathbb{P},\mathbb{Q}) = (2\pi)^{-d/4}\sqrt{\widehat{\psi}(0)}\left\|\widetilde{\Phi} * \mathbb{P} - \widetilde{\Phi} * \mathbb{Q}\right\|_{L^2(\mathbb{R}^d)}. \tag{3.12}$$

The r.h.s. of (3.12) can be interpreted as follows. Let X, Y and N be independent random variables such that $X \sim \mathbb{P}$, $Y \sim \mathbb{Q}$ and $N \sim \widetilde{\Phi}$. This means γ_k is proportional to the L^2-distance computed between the densities associated with the perturbed random variables, $X+N$ and $Y+N$. Note that $\|p-q\|_{L^2(\mathbb{R}^d)}$ is the L^2-distance between the densities of X and Y. Examples of ψ that satisfy the conditions in Theorem 3.4(iii) in addition to the conditions on Φ as mentioned here include the Gaussian and Laplacian kernels on $\mathbb{R}^d$. The result in (3.9) holds even if $\sqrt{\widehat{\psi}} \notin L^1(\mathbb{R}^d)$ as the proof of (iii) can be handled using distribution theory. However, we assumed $\sqrt{\widehat{\psi}} \in L^1(\mathbb{R}^d)$ to keep the proof simple, without delving into distribution theory.

Although we will not be using all the results of Theorem 3.4 in deriving our main results in the following sections, Theorem 3.4 was presented to provide a better intuitive understanding of γ_k. To summarize, the core results of this section are: (a) Proposition 3.2 (combined with Proposition 3.3), which starting from an IPM, derives the embedding in (3.1) through γ_k, and (b) Theorem 3.4*(i)*, which provides an alternative representation for γ_k when k is bounded, continuous and translation invariant on $\mathbb{R}^d$.

3.2 Characteristic Kernels

Having understood how the embedding in (3.1) can be obtained as a special case of IPM, in this section, we address our main question of when is (3.1) injective, i.e., when is γ_k a metric on $M_+^1(\mathcal{X})$. To address this, we first start with the definition of a *characteristic kernel* that was recently introduced in [30].

Definition 3.6 (Characteristic kernel)**.** *A bounded measurable pd kernel k is characteristic to a set $\mathscr{Q}(\mathcal{X}) \subset M^1_+(\mathcal{X})$ of probability measures defined on $\mathcal{X}$ if*

$$\mathbb{P} \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x),\, \mathbb{P} \in \mathscr{Q}(\mathcal{X})$$

is injective, i.e., for $\mathbb{P}, \mathbb{Q} \in \mathscr{Q}(\mathcal{X})$, $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$. k is simply said to be characteristic if it is characteristic to $M^1_+(\mathcal{X})$. The RKHS $\mathcal{H}$ induced by such a k is called a characteristic RKHS.

This means, we are interested in the question of when is k characteristic? Before we get to the characterization of characteristic kernels, the following examples show that there exist bounded measurable kernels that are not characteristic.

Example 3.7 (Trivial kernel)**.** *Let $k(x, y) = \psi(x-y) = C, \forall\, x, y \in \mathbb{R}^d$ with $C > 0$. Using this in (3.5), we have $\gamma^2_k(\mathbb{P}, \mathbb{Q}) = C + C - 2C = 0$ for any $\mathbb{P}, \mathbb{Q} \in M^1_+(\mathcal{X})$, which means k is not characteristic.*

Example 3.8 (Dot product kernel)**.** *Let $k(x, y) = \langle x, y\rangle$, $x, y \in \mathbb{R}^d$. Using this in (3.5), we have*

$$\gamma^2_k(\mathbb{P}, \mathbb{Q}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}}\rangle + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}}\rangle - 2\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}}\rangle = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_2,$$

where $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ represent the means associated with $\mathbb{P}$ and $\mathbb{Q}$ respectively, that is, $\mu_{\mathbb{P}} := \int_{\mathbb{R}^d} x\, d\mathbb{P}(x)$. It is clear that k is not characteristic as $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \nRightarrow \mathbb{P} = \mathbb{Q}$ for all $\mathbb{P}, \mathbb{Q} \in M^1_+(\mathbb{R}^d)$.

Example 3.9 (Polynomial kernel of order 2)**.** *Let $k(x, y) = (1+\langle x, y\rangle)^2$, $x, y \in \mathbb{R}^d$. Using this in (3.6), we have*

$$\begin{aligned}\gamma^2_k(\mathbb{P}, \mathbb{Q}) &= \iint_{\mathbb{R}^d} (1 + 2\langle x, y\rangle + x^T y y^T x)\, d(\mathbb{P} - \mathbb{Q})(x)\, d(\mathbb{P} - \mathbb{Q})(y) \\ &= 2\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_2 + \|\Sigma_{\mathbb{P}} - \Sigma_{\mathbb{Q}} + \mu_{\mathbb{P}}\mu^T_{\mathbb{P}} - \mu_{\mathbb{Q}}\mu^T_{\mathbb{Q}}\|^2_F,\end{aligned}$$

where $\Sigma_{\mathbb{P}}$ and $\Sigma_{\mathbb{Q}}$ represent the covariance matrices associated with $\mathbb{P}$ and $\mathbb{Q}$ respectively, that is, $\Sigma_{\mathbb{P}} := \int_{\mathbb{R}^d} xx^T\, d\mathbb{P}(x) - \mu_{\mathbb{P}}\mu^T_{\mathbb{P}}$. $\|\cdot\|_F$ represents the Frobenius norm. Since $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow (\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$ and $\Sigma_{\mathbb{P}} = \Sigma_{\mathbb{Q}}) \nRightarrow \mathbb{P} = \mathbb{Q}$ for all $\mathbb{P}, \mathbb{Q} \in M^1_+(\mathbb{R}^d)$, k is not characteristic.

Now let us return to the question of when is k characteristic? The following are two characterizations for characteristic kernels (and therefore their corresponding RKHSs) that have already been studied in literature:

1. When $\mathcal{X}$ is a compact metric space, [37] showed that $\mathcal{H}$ is characteristic if k is *universal* in the sense of Steinwart [80, Definition 4], that is, $\mathcal{H}$ is dense in the Banach space of bounded continuous functions with respect to the supremum norm. Examples of such $\mathcal{H}$ include those induced by the Gaussian and Laplacian kernels on every compact subset of $\mathbb{R}^d$.

2. Fukumizu et al. [29,30] extended this characterization to non-compact $\mathcal{X}$ and showed that $\mathcal{H}$ is characteristic if and only if the direct sum of $\mathcal{H}$ and $\mathbb{R}$ is dense in the Banach space of r-integrable (for some $r \geq 1$) functions. Using this characterization, they showed that the RKHSs induced by the Gaussian and Laplacian kernels (supported on the entire $\mathbb{R}^d$) are characteristic.

In the following sections, we provide alternative conditions for characteristic RKHSs which address several limitations of the foregoing. First, it can be difficult to verify the conditions of denseness in both of the above characterizations. Second, universality is in any case an overly restrictive condition because universal kernels assume $\mathcal{X}$ to be compact, that is, they induce a metric only on the space of probability measures that are supported on compact $\mathcal{X}$.

In Section 3.2.1, we present the simple characterization that integrally strictly pd kernels (see Section 2.2 for the definition) are characteristic, that is, the induced RKHS is characteristic. This condition is more natural—strict pd is a natural property of interest for kernels, unlike the denseness condition—and much easier to understand than the characterizations mentioned above. Examples of integrally strictly pd kernels on $\mathbb{R}^d$ include the Gaussian, Laplacian, inverse multiquadratics, Matérn kernel family, B_{2n+1}-splines, etc.

Although the above characterization of integrally strictly pd kernels being characteristic is simple to understand, it is only a sufficient condition and does not provide an answer for kernels that are not integrally strictly pd,[3] for exam-

[3] Proposition 2.6 shows that integrally strictly pd kernels are strictly pd. Therefore, examples of kernels that are not integrally strictly pd include those kernels that are not strictly pd.

ple, a Dirichlet kernel. Therefore, in Section 3.2.2, we provide an easily checkable condition, after making some assumptions on the kernel. We present a complete characterization of characteristic kernels when the kernel is translation invariant on $\mathbb{R}^d$. We show that a bounded continuous translation invariant kernel on $\mathbb{R}^d$ is characteristic if and only if the support of the Fourier transform of the kernel is the entire $\mathbb{R}^d$. This condition is easy to check compared to the characterizations described above. We also show that all compactly supported translation invariant kernels on $\mathbb{R}^d$ are characteristic. Note, however, that the characterization of integral strict positive definiteness in Section 3.2.1 does not assume $\mathcal{X}$ to be $\mathbb{R}^d$ nor k to be translation invariant.

We extend the result of Section 3.2.2 to $\mathcal{X}$ being a d-Torus, that is, $\mathbb{T}^d = S^1 \times \overset{d}{\ldots} \times S^1 \equiv [0, 2\pi)^d$, where S^1 is a circle. In Section 3.2.3, we show that a translation invariant kernel on $\mathbb{T}^d$ is characteristic if and only if the Fourier series coefficients of the kernel are positive, that is, the support of the Fourier spectrum is the entire $\mathbb{Z}^d$. The proof of this result is similar in flavor to the one in Section 3.2.2. As examples, the Poisson kernel can be shown to be characteristic, while the Dirichlet kernel is not.

The main results of this section are summarized in Table 3.1.

3.2.1 Integrally Strictly Positive Definite Kernels are Characteristic

Compared to the existing characterizations in literature [29,30,37], the following result provides a more natural and easily understandable characterization for characteristic kernels, namely that integrally strictly pd kernels are characteristic to $M_+^1(\mathcal{X})$.

Theorem 3.10 (Integrally strictly pd kernels are characteristic)**.** *If k is an integrally strictly pd kernel on $\mathcal{X}$, then it is characteristic to $M_+^1(\mathcal{X})$.*

Before proving Theorem 3.10, we provide a supplementary result in Lemma 3.11 that provides necessary and sufficient conditions for a kernel *not* to be characteristic. We show that choosing k to be integrally strictly pd violates the conditions in Lemma 3.11, and k is therefore characteristic to $M_+^1(\mathcal{X})$.

Lemma 3.11. *Let k be measurable and bounded on $\mathcal{X}$. Then $\exists\,\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ if and only if there exists $\mu \in M_b(\mathcal{X})\backslash\{0\}$ that satisfies:*

(i) $\iint_{\mathcal{X}} k(x, y)\, d\mu(x)\, d\mu(y) = 0,$

(ii) $\mu(\mathcal{X}) = 0.$

Proof. ($\Leftarrow$) Suppose there exists $\mu \in M_b(\mathcal{X})\backslash\{0\}$ that satisfies *(i)* and *(ii)* in Lemma 3.11. By the Jordan decomposition theorem [26, Theorem 3.4], there exist unique positive measures μ^+ and μ^- such that $\mu = \mu^+ - \mu^-$ and $\mu^+ \perp \mu^-$ (μ^+ and μ^- are singular). By *(ii)*, we have $\mu^+(\mathcal{X}) = \mu^-(\mathcal{X}) =: \alpha$. Define $\mathbb{P} = \alpha^{-1}\mu^+$ and $\mathbb{Q} = \alpha^{-1}\mu^-$. Clearly, $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$. Then, by (3.6), we have

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \iint_{\mathcal{X}} k(x, y)\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y) = \alpha^{-2} \iint_{\mathcal{X}} k(x, y)\, d\mu(x)\, d\mu(y) \overset{(a)}{=} 0,$$

where *(a)* is obtained by invoking *(i)*. So, we have constructed $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$.

($\Rightarrow$) Suppose $\exists\,\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. Let $\mu = \mathbb{P} - \mathbb{Q}$. Clearly $\mu \in M_b(\mathcal{X})\backslash\{0\}$. Note that by (3.6),

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \iint_{\mathcal{X}} k(x, y)\, d(\mathbb{P} - \mathbb{Q})(x)\, d(\mathbb{P} - \mathbb{Q})(y) = \iint_{\mathcal{X}} k(x, y)\, d\mu(x)\, d\mu(y),$$

and therefore *(i)* follows. □

Proof of Theorem 3.10. Since k is integrally strictly pd on $\mathcal{X}$, it satisfies (2.6), which means there does not exist $\mu \in M_b(\mathcal{X})\backslash\{0\}$ that satisfies *(i)* in Lemma 3.11. Therefore, by Lemma 3.11, there does not exist $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$, which implies k is characteristic. □

Examples of integrally strictly pd kernels on $\mathbb{R}^d$ include the Gaussian, Laplacian, inverse multiquadratics, etc., which are translation invariant kernels on $\mathbb{R}^d$. A *translation variant* integrally strictly pd kernel, $\widetilde{k}$, can be obtained from a translation invariant integrally strictly pd kernel, k, as $\widetilde{k}(x, y) = f(x)k(x, y)f(y)$, where

$f : \mathcal{X} \to \mathbb{R}$ is a bounded continuous function. A simple example of a translation variant integrally strictly pd kernel on $\mathbb{R}^d$ is $\widetilde{k}(x,y) = \exp(\sigma\langle x,y\rangle)$, $\sigma > 0$, where we have chosen $f(\cdot) = \exp(\sigma\|\cdot\|_2^2/2)$ and $k(x,y) = \exp(-\sigma\|x-y\|_2^2/2)$, $\sigma > 0$. Clearly, this kernel is characteristic on compact subsets of $\mathbb{R}^d$. The same result can also be obtained from the fact that $\widetilde{k}$ is universal on compact subsets of $\mathbb{R}^d$ [80, Section 3, Example 1], recalling that universal kernels are characteristic [37, Theorem 3].

Although the condition for characteristic k in Theorem 3.10 is easy to understand compared to other characterizations in literature, it is not always easy to check for integral strict positive definiteness of k. In the following section, we assume $\mathcal{X} = \mathbb{R}^d$ and k to be translation invariant and present a complete characterization for characteristic k which is simple to check.

3.2.2 Characterization for Translation Invariant Kernels on $\mathbb{R}^d$

The complete, detailed proofs of the main results in this section are provided in Section 3.2.4. Throughout this section, we assume as follows.

Assumption 3.12. *$k(x,y) = \psi(x-y)$ where ψ is a bounded continuous real-valued positive definite function on $\mathcal{X} = \mathbb{R}^d$.*

The following theorem characterizes all translation invariant kernels in $\mathbb{R}^d$ that are characteristic.

Theorem 3.13. *Suppose k satisfies Assumption 3.12. Then k is characteristic if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$, where Λ is defined as in (2.2).*

First, note that the condition $\text{supp}(\Lambda) = \mathbb{R}^d$ is easy to check compared to all other, aforementioned characterizations for characteristic k. Although the Gaussian and Laplacian kernels are shown to be characteristic by all the characterizations we have mentioned so far, the case of B_{2n+1}-splines is addressed only by Theorem 3.13, which shows them to be characteristic (note that B_{2n+1}-splines being integrally strictly pd also follows from Theorem 3.13). In fact, one can provide a more general result on compactly supported translation invariant kernels, which we do later in Corollary 3.14. By Theorem 3.13, the sinc kernel is not characteristic, which is not easy to show using other characterizations. By combining Theorem 3.10 with Theorem 3.13, it can be shown that the sinc, Poisson, Dirichlet and Féjer kernels are not integrally strictly pd. Therefore, for translation invariant kernels on $\mathbb{R}^d$, the integral strict positive definiteness of the kernel (or the lack of it) can be tested using Theorems 3.10 and 3.13.

Proof of Theorem 3.13. We provide an outline of the complete proof, which is presented in Section 3.2.4. The sufficient condition in Theorem 3.13 is simple to prove and follows from Theorem 3.4*(i)*, whereas we need a supplementary result to prove its necessity, which is presented in Lemma 3.21 (see Section 3.2.4). Proving the necessity of Theorem 3.13 is equivalent to showing that if $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, then $\exists\, \mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in M^1_+(\mathcal{X})$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. In Lemma 3.21, we present equivalent conditions for the existence of $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ if $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, using which we prove the necessity of Theorem 3.13. □

The whole family of compactly supported translation invariant continuous bounded kernels on $\mathbb{R}^d$ is characteristic, as shown by the following corollary to Theorem 3.13.

Corollary 3.14. *Suppose $k \neq 0$ satisfies Assumption 3.12 and* $\text{supp}(\psi)$ *is compact. Then k is characteristic.*

Proof. Since $\psi \in C_b(\mathbb{R}^d)$ is compactly supported on $\mathbb{R}^d$, by (C.4), $\psi \in \mathcal{D}'_d$. Therefore, by the Paley-Wiener theorem (Theorem C.9), $\widehat{\psi}$ is the restriction to $\mathbb{R}^d$ of an *entire function*[4] on $\mathbb{C}^d$, which means $\widehat{\psi}$ is an analytic function on $\mathbb{R}^d$. Suppose $\text{supp}(\widehat{\psi})$ is compact, which means there exists an open set, $U \subset \mathbb{R}^d$ such that $\widehat{\psi}(x) = 0, \forall\, x \in U$. But being analytic, this implies that $\widehat{\psi}(x) = 0, \forall\, x \in \mathbb{R}^d$, that is, $\psi = 0$, which leads to a contradiction. Therefore, $\widehat{\psi}$ cannot be compactly supported, that is, $\text{supp}(\widehat{\psi}) = \mathbb{R}^d$, and the result follows from Theorem 3.13. □

[4] Let $D \subset \mathbb{C}^d$ be an open subset and $f : D \to \mathbb{C}$ be a function. f is said to be *holomorphic* (or *analytic*) at the point $z_0 \in D$ if $f'(z_0) := \lim_{z \to z_0} \frac{f(z_0)-f(z)}{z_0 - z}$ exists. Moreover, f is called holomorphic if it is holomorphic at every $z_0 \in D$. f is called an *entire function* if f is holomorphic and $D = \mathbb{C}^d$.

The above result is interesting in practice because of the computational advantage in dealing with compactly supported kernels. Note that proving such a general result for compactly supported kernels on $\mathbb{R}^d$ is not straightforward (maybe not even possible) with the other characterizations.

As a corollary to Theorem 3.13, the following result provides a method to construct new characteristic kernels from a given one.

Corollary 3.15. *Let k, k_1 and k_2 satisfy Assumption 3.12. Suppose k is characteristic and $k_2 \neq 0$. Then $k + k_1$ and $k \cdot k_2$ are characteristic.*

Proof. Since k, k_1 and k_2 satisfy Assumption 3.12, $k + k_1$ and $k_2 \cdot k$ also satisfy Assumption 3.12. In addition,

$$\begin{aligned}
(k+k_1)(x,y) &:= k(x,y) + k_1(x,y) \\
&= \psi(x-y) + \psi_1(x-y) \\
&= \int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle}\, d(\Lambda+\Lambda_1)(\omega), \\
(k \cdot k_2)(x,y) &:= k(x,y)k_2(x,y) \\
&= \psi(x-y)\psi_2(x-y) \\
&= \iint_{\mathbb{R}^d} e^{-i\langle x-y,\omega+\xi\rangle}\, d\Lambda(\omega)\, d\Lambda_2(\xi) \\
&\overset{(a)}{=}: \int_{\mathbb{R}^d} e^{-i\langle x-y,\omega\rangle}\, d(\Lambda * \Lambda_2)(\omega),
\end{aligned}$$

where (a) follows from the definition of convolution of measures (see [65, Section 9.14] for details). Since k is characteristic, that is, $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, and $\mathrm{supp}(\Lambda) \subset \mathrm{supp}(\Lambda + \Lambda_1)$, we have $\mathrm{supp}(\Lambda+\Lambda_1) = \mathbb{R}^d$ and therefore $k + k_1$ is characteristic. Similarly, since $\mathrm{supp}(\Lambda) \subset \mathrm{supp}(\Lambda * \Lambda_2)$, we have $\mathrm{supp}(\Lambda * \Lambda_2) = \mathbb{R}^d$ and therefore, $k \cdot k_2$ is characteristic. □

Note that in the above result, we do not need k_1 or k_2 to be characteristic. Therefore, one can generate all sorts of kernels that are characteristic by starting with a characteristic kernel, k.

Since a radial kernel in (2.5) defined on $\mathbb{R}^d$ is also translation invariant (see Proposition 2.3), we have the following corollary to Theorem 3.13 that provides a necessary and sufficient condition for it to be characteristic.

Corollary 3.16. *Suppose k is radial on $\mathbb{R}^d$, i.e.,*

$$k(x,y) = \int_{[0,\infty)} e^{-t\|x-y\|_2^2}\, d\nu(t),\; x,y \in \mathbb{R}^d,$$

where $\nu \in M_b^+([0,\infty))$. Then k is characteristic if and only if $\operatorname{supp}(\nu) \neq \{0\}$.

Proof. From the proof of Proposition 2.3, we have $k(x,y) = \psi(x-y)$, where $\psi(x) = \int_{\mathbb{R}^d} e^{-i\langle x,\omega\rangle}\, d\Lambda(\omega)$, $d\Lambda(\omega) = \phi(\omega)\, d\omega$ and

$$\phi(\omega) = \int_{[0,\infty)} \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|\omega\|_2^2}{4t}}\, d\nu(t).$$

Suppose $\operatorname{supp}(\nu) \neq \{0\}$, which means $\operatorname{supp}(\phi) = \operatorname{supp}(\Lambda) = \mathbb{R}^d$ and therefore k is characteristic by Theorem 3.13. On the other hand, if $\operatorname{supp}(\nu) = \{0\}$, then $k(x,y) = 1$, which by Example 3.7 is not characteristic. □

So far, we have considered characterizations for k such that it is characteristic to $M_+^1(\mathbb{R}^d)$. We showed in Theorem 3.13 that kernels with $\operatorname{supp}(\Lambda) \subsetneq \mathbb{R}^d$ are not characteristic to $M_+^1(\mathbb{R}^d)$. Now, we can question whether such kernels can be characteristic to some proper subset $\mathscr{Q}(\mathbb{R}^d)$ of $M_+^1(\mathbb{R}^d)$. The following result addresses this. Note that these kernels, that is, the kernels with $\operatorname{supp}(\Lambda) \subsetneq \mathbb{R}^d$ are usually not useful in practice, especially in statistical inference applications, because the conditions on $\mathscr{Q}(\mathbb{R}^d)$ are usually not easy to check. On the other hand, the following result is of theoretical interest: along with Theorem 3.13, it completes the characterization of characteristic kernels that are translation invariant on $\mathbb{R}^d$. Before we state the result, we denote $\mathbb{P} \ll \mathbb{Q}$ to mean that $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$.

Theorem 3.17. *Let $\mathscr{P}_1(\mathbb{R}^d) := \{\mathbb{P} \in M_+^1(\mathbb{R}^d) : \phi_{\mathbb{P}} \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d),\ \mathbb{P} \ll \lambda$ and $\operatorname{supp}(\mathbb{P})$ is compact$\}$, where λ is the Lebesgue measure. Suppose k satisfies Assumption 3.12 and $\operatorname{supp}(\Lambda) \subsetneq \mathbb{R}^d$ has a non-empty interior, where Λ is defined as in (2.2). Then k is characteristic to $\mathscr{P}_1(\mathbb{R}^d)$.*

Proof. See Section 3.2.4. □

Although, by Theorem 3.13, the kernels with $\operatorname{supp}(\Lambda) \subsetneq \mathbb{R}^d$ are not characteristic to $M_+^1(\mathbb{R}^d)$, Theorem 3.17 shows that there exists a subset of $M_+^1(\mathbb{R}^d)$

to which a subset of these kernels are characteristic. This type of result is not available for the previously mentioned characterizations. An example of a kernel that satisfies the conditions in Theorem 3.17 is the sinc kernel (see Example 2.4) which has $\text{supp}(\Lambda) = [-\sigma, \sigma]^d$. The condition that $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ has a non-empty interior is important for Theorem 3.17 to hold. If $\text{supp}(\Lambda)$ has an empty interior (examples include periodic kernels), then one can construct $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_1(\mathbb{R}^d)$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. This is illustrated in Example 3.24 of Section 3.2.4.

So far, we have characterized the characteristic property of kernels that satisfy (a) $\text{supp}(\Lambda) = \mathbb{R}^d$ or (b) $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ with $\text{int}(\text{supp}(\Lambda)) \neq \emptyset$. In the following section, we investigate kernels that have $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ with $\text{int}(\text{supp}(\Lambda)) = \emptyset$, examples of which include periodic kernels on $\mathbb{R}^d$. This discussion uses the fact that a periodic function on $\mathbb{R}^d$ can be treated as a function on $\mathbb{T}^d$, the d-Torus.

3.2.3 Characterization for Translation Invariant Kernels on $\mathbb{T}^d$

Let $\mathscr{X} = \times_{j=1}^d [0, \tau_j)$ and $\tau := (\tau_1, \ldots, \tau_d)$. A function defined on $\mathscr{X}$ with periodic boundary conditions is equivalent to considering a periodic function on $\mathbb{R}^d$ with period τ. With no loss of generality, we can choose $\tau_j = 2\pi$, $\forall\, j$ which yields $\mathscr{X} = [0, 2\pi)^d =: \mathbb{T}^d$, called the d-Torus. The results presented here hold for any $0 < \tau_j < \infty$, $\forall\, j$ but we choose $\tau_j = 2\pi$ for simplicity. Similar to Assumption 3.12, we make the following assumption.

Assumption 3.18. *$k(x, y) = \psi((x-y)_{mod\ 2\pi})$, where ψ is a continuous real-valued positive definite function on $\mathscr{X} = \mathbb{T}^d$.*

We now state the result that defines characteristic kernels on $\mathbb{T}^d$.

Theorem 3.19. *Suppose k satisfies Assumption 3.18. Then k is characteristic to $M_+^1(\mathbb{T}^d)$ if and only if $A_\psi(0) \geq 0$, $A_\psi(n) > 0, \forall\, n \neq 0$, where A_ψ is defined in (2.4).*

The proof is provided in Section 3.2.4 and the idea is similar to that of Theorem 3.13. Based on the above result, one can generate characteristic kernels by constructing an infinite sequence of positive numbers that are summable and then using them in (2.4). Note that the Poisson kernel on $\mathbb{T}$ is characteristic while the Dirichlet and Féjer kernels are not. Some other examples of characteristic kernels on $\mathbb{T}$ are:

(1) $k(x,y) = e^{\alpha \cos(x-y)} \cos(\alpha \sin(x-y))$, $0 < \alpha \leq 1 \leftrightarrow A_\psi(0) = 1$, $A_\psi(n) = \frac{\alpha^{|n|}}{2|n|!}$, $\forall\, n \neq 0$.

(2) $k(x,y) = -\log(1 - 2\alpha \cos(x-y) + \alpha^2)$, $0 < \alpha < 1 \leftrightarrow A_\psi(0) = 0$, $A_\psi(n) = \frac{\alpha^{|n|}}{|n|}$, $\forall\, n \neq 0$.

(3) $k(x,y) = (\pi - (x-y)_{mod\ 2\pi})^2 \leftrightarrow A_\psi(0) = \frac{\pi^2}{3}$, $A_\psi(n) = \frac{2}{n^2}$, $\forall\, n \neq 0$.

(4) $k(x,y) = \frac{\sinh \alpha}{\cosh \alpha - \cos(x-y)}$, $\alpha > 0 \leftrightarrow A_\psi(0) = 1, A_\psi(n) = e^{-\alpha|n|}$, $\forall\, n \neq 0$.

(5) $k(x,y) = \frac{\pi \cosh(\alpha(\pi - (x-y)_{mod\, 2\pi}))}{\alpha \sinh(\pi\alpha)} \leftrightarrow A_\psi(0) = \frac{1}{\alpha^2}$, $A_\psi(n) = \frac{1}{n^2+\alpha^2}$, $\forall\, n \neq 0$.

The following result relates characteristic kernels and universal kernels defined on $\mathbb{T}^d$.

Corollary 3.20. *Let k be a characteristic kernel satisfying Assumption 3.18 with $A_\psi(0) > 0$. Then k is also universal.*

Proof. Since k is characteristic with $A_\psi(0) > 0$, we have $A_\psi(n) > 0$, $\forall\, n$. Therefore, by Corollary 11 of [80], k is universal. □

Since k being universal implies that it is characteristic, the above result shows that the converse is not true (though almost true except that $A_\psi(0)$ can be zero for characteristic kernels). The condition on A_ψ in Theorem 3.19, that is, $A_\psi(0) \geq 0$, $A_\psi(n) > 0$, $\forall\, n \neq 0$ can be equivalently written as $\text{supp}(A_\psi) = \mathbb{Z}^d$ or $\text{supp}(A_\psi) = \mathbb{Z}^d \backslash \{0\}$. Therefore, Theorems 3.13 and 3.19 are of similar flavor. In fact, these results can be generalized to locally compact Abelian groups. Fukumizu et al. [31] show that a bounded continuous translation invariant kernel on a locally compact Abelian group G is characteristic to the set of all probability measures on G if and only if the support of the Fourier transform of the translation invariant kernel is the dual group of G. In our case, $(\mathbb{R}^d, +)$ and $(\mathbb{T}^d, +)$ are locally compact

Abelian groups with $(\mathbb{R}^d, +)$ and $(\mathbb{Z}^d, +)$ as their respective dual groups. In [31], these results are also extended to translation invariant kernels on non-Abelian compact groups and the semigroup $\mathbb{R}^d_+$.

3.2.4 Proofs

First, we present a supplementary result in Lemma 3.21 that will be used to prove Theorem 3.13. The idea of Lemma 3.21 is to characterize the equivalent conditions for the existence of $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ when $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Its proof relies on the properties of characteristic functions, which we have collected in Theorem C.7.

Lemma 3.21. *Let $\mathscr{P}_0(\mathbb{R}^d) := \{\mathbb{P} \in M^1_+(\mathbb{R}^d) : \phi_\mathbb{P} \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$ and $\mathbb{P} \ll \lambda\}$, where λ is the Lebesgue measure. Suppose k satisfies Assumption 3.12 and* $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$*, where Λ is defined as in (2.2). Then, for any $\mathbb{Q} \in \mathscr{P}_0(\mathbb{R}^d)$, $\exists\, \mathbb{P} \neq \mathbb{Q}$, $\mathbb{P} \in \mathscr{P}_0(\mathbb{R}^d)$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ if and only if there exists a non-zero function $\theta : \mathbb{R}^d \to \mathbb{C}$ that satisfies the following conditions:*

(i) $\theta \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$ *is conjugate symmetric, that is,* $\overline{\theta(x)} = \theta(-x)$, $\forall\, x \in \mathbb{R}^d$,

(ii) $\theta^\vee \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$,

(iii) $\int_{\mathbb{R}^d} |\theta(x)|^2 \, d\Lambda(x) = 0$,

(iv) $\theta(0) = 0$,

(v) $\inf_{x \in \mathbb{R}^d} \{\theta^\vee(x) + q(x)\} \geq 0$.

Proof. Define $L^1 := L^1(\mathbb{R}^d)$, $L^2 := L^2(\mathbb{R}^d)$ and $C_b := C_b(\mathbb{R}^d)$.
$(\Leftarrow)$ Suppose there exists a non-zero function θ satisfying *(i)* – *(v)*. For any $\mathbb{Q} \in \mathscr{P}_0(\mathbb{R}^d)$, we have $\phi_\mathbb{Q} \in L^1 \cup L^2$ and $\phi_\mathbb{Q} \in C_b$ (by Theorem C.7), that is, $\phi_\mathbb{Q} \in (L^1 \cup L^2) \cap C_b$. Now, consider the case of $\phi_\mathbb{Q} \in L^1 \cap C_b$. Since $\phi_\mathbb{Q} \in L^1$, by the inversion theorem for characteristic functions (see [23, Theorem 9.5.4]), $\mathbb{Q}$ is absolutely continuous w.r.t. λ. If q is the Radon-Nikodym derivative of $\mathbb{Q}$ w.r.t. λ, then $q = [\overline{\phi_\mathbb{Q}}]^\vee \in L^1$. In addition, by the Riemann-Lebesgue lemma

(Lemma C.8), we have $q \in C_0(\mathbb{R}^d) \subset C_b$, which therefore implies $q \in L^1 \cap C_b$. When $\phi_{\mathbb{Q}} \in L^2 \cap C_b$, the Fourier transform in the L^2 sense (see Section C.1.2) implies that $q = [\overline{\phi_{\mathbb{Q}}}]^\vee \in L^1 \cap L^2$. Therefore, $q \in L^1 \cap (L^2 \cup C_b)$. Define $p := q + \theta^\vee$. Clearly $p \in L^1 \cap (L^2 \cup C_b)$. In addition, $\overline{\phi_{\mathbb{P}}} = \widehat{p} = \widehat{q} + \widehat{\theta^\vee} = \overline{\phi_{\mathbb{Q}}} + \theta \in (L^1 \cup L^2) \cap C_b$. Since θ is conjugate symmetric, $\theta^\vee$ is real valued and so is p. Consider

$$\int_{\mathbb{R}^d} p(x)\, dx = \int_{\mathbb{R}^d} q(x)\, dx + \int_{\mathbb{R}^d} \theta^\vee(x)\, dx = 1 + \theta(0) = 1.$$

(v) implies that p is nonnegative. Therefore, p is the Radon-Nikodym derivative of a probability measure $\mathbb{P}$ w.r.t. λ, where $\mathbb{P}$ is such that $\mathbb{P} \neq \mathbb{Q}$ and $\mathbb{P} \in \mathscr{P}_0(\mathbb{R}^d)$. By (3.7), we have

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} |\phi_{\mathbb{P}}(x) - \phi_{\mathbb{Q}}(x)|^2\, d\Lambda(x) = \int_{\mathbb{R}^d} |\theta(x)|^2\, d\Lambda(x) = 0.$$

($\Rightarrow$) Suppose that there exists $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_0(\mathbb{R}^d)$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. Define $\theta := \phi_{\mathbb{P}} - \phi_{\mathbb{Q}}$. We need to show that θ satisfies *(i) – (v)*. Recalling Theorem C.7, $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_0(\mathbb{R}^d)$ implies $\phi_{\mathbb{P}}, \phi_{\mathbb{Q}} \in (L^1 \cup L^2) \cap C_b$ and $p, q \in L^1 \cap (L^2 \cup C_b)$. Therefore, $\theta = \overline{\phi}_{\mathbb{P}} - \overline{\phi}_{\mathbb{Q}} \in (L^1 \cup L^2) \cap C_b$ and $\theta^\vee = p - q \in L^1 \cap (L^2 \cup C_b)$. By Theorem C.7, $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ are conjugate symmetric and so is θ. Therefore θ satisfies *(i)* and $\theta^\vee$ satisfies *(ii)*. θ satisfies *(iv)* as

$$\theta(0) = \int_{\mathbb{R}^d} \theta^\vee(x)\, dx = \int_{\mathbb{R}^d} (p(x) - q(x))\, dx = 0.$$

Non-negativity of p yields *(v)*. By (3.7), $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ implies *(iii)*. □

Remark 3.22. *Note that the dependence of θ on the kernel appears in the form of (iii) in Lemma 3.21. This condition shows that $\lambda(\mathrm{supp}(\theta) \cap \mathrm{supp}(\Lambda)) = 0$, that is, the supports of θ and Λ are disjoint w.r.t. the Lebesgue measure, λ. In other words, $\mathrm{supp}(\theta) \subset \mathrm{cl}(\mathbb{R}^d \backslash \mathrm{supp}(\Lambda))$. So, the idea is to introduce the perturbation, θ over an open set, U where $\Lambda(U) = 0$. The remaining conditions characterize the nature of this perturbation so that the constructed measure, $p = q + \theta^\vee$, is a valid probability measure. Conditions (i), (ii) and (iv) simply follow from $\theta = \phi_{\mathbb{P}} - \phi_{\mathbb{Q}}$, while (v) ensures that $p(x) \geq 0$, $\forall\, x$.*

Using Lemma 3.21, we now present the proof of Theorem 3.13.

Proof of Theorem 3.13. The sufficiency follows from (3.7): if $\text{supp}(\Lambda)=\mathbb{R}^d$, then $\gamma_k^2(\mathbb{P},\mathbb{Q})=\int_{\mathbb{R}^d}|\phi_{\mathbb{P}}(x)-\phi_{\mathbb{Q}}(x)|^2\,d\Lambda(x)=0\Rightarrow\phi_{\mathbb{P}}=\phi_{\mathbb{Q}}$, a.e. Recalling from Theorem C.7 that $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ are uniformly continuous on $\mathbb{R}^d$, we have that $\mathbb{P}=\mathbb{Q}$, and therefore k is characteristic. To prove necessity, we need to show that if $\text{supp}(\Lambda)\subsetneq\mathbb{R}^d$, then there exists $\mathbb{P}\neq\mathbb{Q}$, $\mathbb{P},\mathbb{Q}\in M_+^1(\mathbb{R}^d)$ such that $\gamma_k(\mathbb{P},\mathbb{Q})=0$. By Lemma 3.21, this is equivalent to showing that there exists a non-zero θ satisfying the conditions in Lemma 3.21. Below, we provide a constructive procedure for such a θ when $\text{supp}(\Lambda)\subsetneq\mathbb{R}^d$, thereby proving the result.

Consider the following function, $f_{\beta,\omega_0}\in C^\infty(\mathbb{R}^d)$ supported in $[\omega_0-\beta,\omega_0+\beta]$,

$$f_{\beta,\omega_0}(\omega)=\prod_{j=1}^d h_{\beta_j,\omega_{0,j}}(\omega_j)\text{ with }h_{a,b}(y):=\mathbb{1}_{[-a,a]}(y-b)\,e^{-\frac{a^2}{a^2-(y-b)^2}},$$

where $\omega=(\omega_1,\ldots,\omega_d)$, $\omega_0=(\omega_{0,1},\ldots,\omega_{0,d})$, $\beta=(\beta_1,\ldots,\beta_d)$, $a\in\mathbb{R}_{++}$, $b\in\mathbb{R}$ and $y\in\mathbb{R}$. Since $\text{supp}(\Lambda)\subsetneq\mathbb{R}^d$, there exists an open set $U\subset\mathbb{R}^d$ such that $\Lambda(U)=0$. So, there exists $\beta\in\mathbb{R}^d_{++}$ and $\omega_0>\beta$ (element-wise inequality) such that $[\omega_0-\beta,\omega_0+\beta]\subset U$. Let

$$\theta=\alpha(f_{\beta,\omega_0}+f_{\beta,-\omega_0}),\ \alpha\in\mathbb{R}\backslash\{0\},$$

which implies $\text{supp}(\theta)=[-\omega_0-\beta,-\omega_0+\beta]\cup[\omega_0-\beta,\omega_0+\beta]$ is compact. Clearly $\theta\in\mathcal{D}_d\subset\mathcal{S}_d$ which implies $\theta^\vee\in\mathcal{S}_d\subset L^1(\mathbb{R}^d)\cap L^2(\mathbb{R}^d)$. Therefore, by construction, θ satisfies *(i)* – *(iv)* in Lemma 3.21. Since $\int_{\mathbb{R}^d}\theta^\vee(x)\,dx=\theta(0)=0$ (by construction), $\theta^\vee$ will take negative values, so we need to show that there exists $\mathbb{Q}\in\mathscr{P}_0(\mathbb{R}^d)$ such that *(v)* in Lemma 3.21 holds. Let $\mathbb{Q}$ be such that it has a density given by

$$q(x)=C_l\prod_{j=1}^d\frac{1}{(1+|x_j|^2)^l},\ l\in\mathbb{N}\text{ where }\ C_l=\prod_{j=1}^d\left(\int_{\mathbb{R}}(1+|x_j|^2)^{-l}\,dx_j\right)^{-1},$$

and $x=(x_1,\ldots,x_d)$. It can be verified that choosing α such that

$$0<|\alpha|\leq\frac{C_l}{2\sup_x\left|\prod_{j=1}^d h^\vee_{\beta_j,0}(x_j)(1+|x_j|^2)^l\cos(\langle\omega_0,x\rangle)\right|}<\infty,$$

ensures that θ satisfies *(v)* in Lemma 3.21. The existence of finite α is guaranteed as $h_{a,0}\in\mathcal{D}_1\subset\mathcal{S}_1$ which implies $h^\vee_{a,0}\in\mathcal{S}_1$, $\forall\,a$. We conclude there exists a non-zero θ as claimed earlier, which completes the proof. □

To elucidate the necessity part in the above proof, in the following, we present a simple example that provides an intuitive understanding about the construction of θ such that for a given $\mathbb{Q}$, $\mathbb{P} \neq \mathbb{Q}$ can be constructed with $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$.

Example 3.23. *Let $\mathbb{Q}$ be a Cauchy distribution in $\mathbb{R}$, that is, $q(x) = \frac{1}{\pi(1+x^2)}$ with characteristic function, $\phi_{\mathbb{Q}}(\omega) = \frac{1}{\sqrt{2\pi}} e^{-|\omega|}$ in $L^1(\mathbb{R})$. Let ψ be a sinc kernel, that is, $\psi(x) = \sqrt{\frac{2}{\pi}} \frac{\sin(\beta x)}{x}$ with Fourier transform*[5] *given by $\widehat{\psi}(\omega) = \mathbb{1}_{[-\beta,\beta]}(\omega)$ and $\text{supp}(\widehat{\psi}) = [-\beta,\beta] \subsetneq \mathbb{R}$. Let θ be*

$$\theta(\omega) = \frac{\alpha}{2i}\left[*_1^N \mathbb{1}_{\left[-\frac{\beta}{2},\frac{\beta}{2}\right]}(\omega)\right] * \left[\delta(\omega-\omega_0) - \delta(\omega+\omega_0)\right],$$

where $|\omega_0| \geq \left(\frac{N+2}{2}\right)\beta$, $N \geq 2$ and $\alpha \neq 0$. $*_1^N$ represents the N-fold convolution. Note that θ is such that $\text{supp}(\theta) \cap \text{supp}(\widehat{\psi})$ is a null set w.r.t. the Lebesgue measure, which satisfies (iii) in Lemma 3.21. It is easy to verify that $\theta \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ also satisfies conditions (i) and (iv) in Lemma 3.21. $\theta^\vee$ can be computed as

$$\theta^\vee(x) = \frac{2^N \alpha}{\sqrt{2\pi}} \sin(\omega_0 x) \frac{\sin^N\left(\frac{\beta x}{2}\right)}{x^N},$$

and $\theta^\vee \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ satisfies (ii) in Lemma 3.21. Choose

$$0 < |\alpha| \leq \frac{\sqrt{2}}{\sqrt{\pi}\beta^N \sup_x \left|(1+x^2)\sin(\omega_0 x)\text{sinc}^N\left(\frac{\beta x}{2\pi}\right)\right|},$$

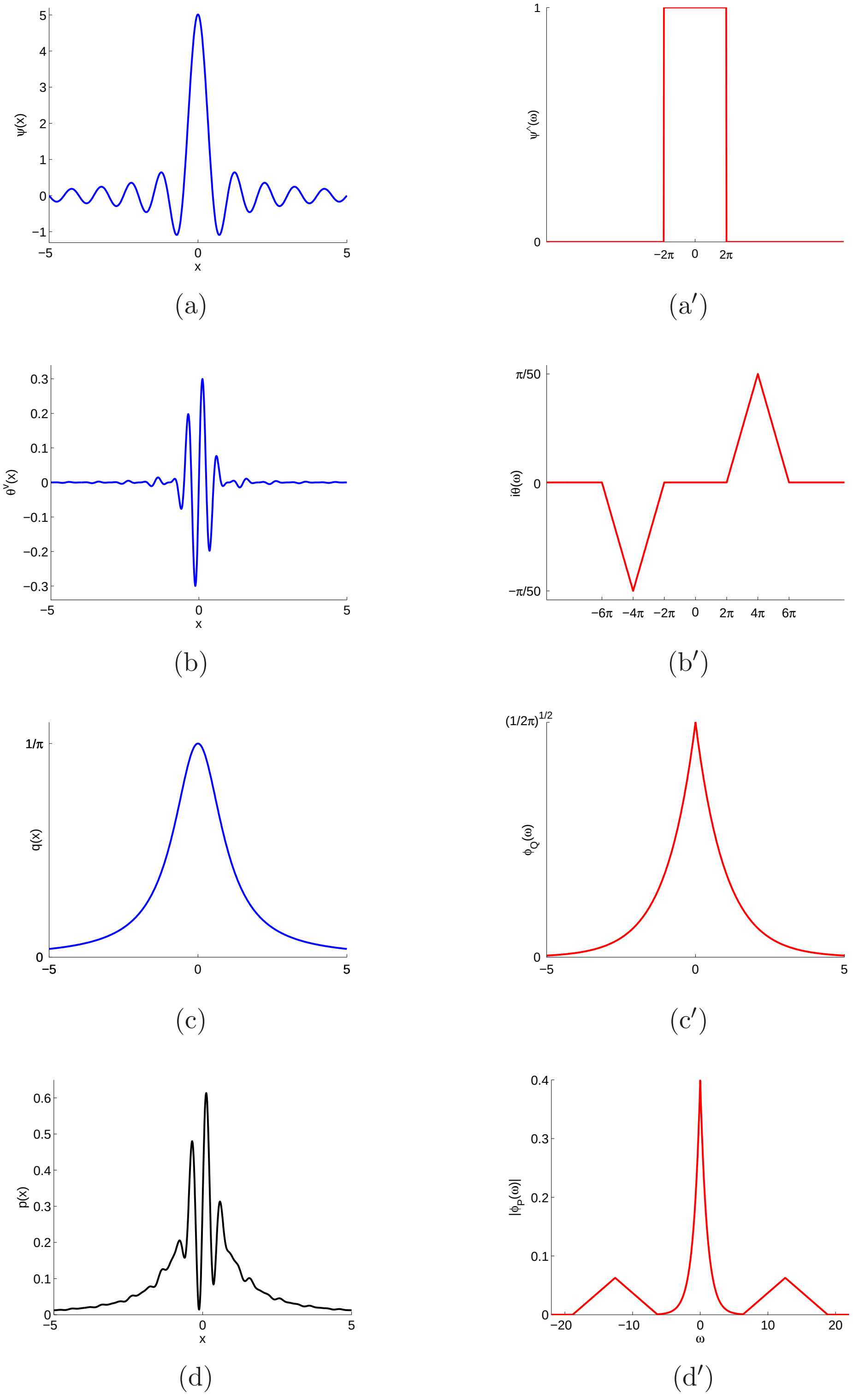
where $\text{sinc}(x) := \frac{\sin(\pi x)}{\pi x}$. Define $g(x) := \sin(\omega_0 x)\text{sinc}^N\left(\frac{\beta x}{2\pi}\right)$. Since $g \in \mathcal{S}_1$, $0 < \sup_x |(1+x^2)g(x)| < \infty$ and, therefore, α is a finite non-zero number. It is easy to see that θ satisfies (v) of Lemma 3.21. Then, by Lemma 3.21, there exists $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P} \in \mathscr{P}_0(\mathbb{R}^d)$, given by

$$p(x) = \frac{1}{\pi(1+x^2)} + \frac{2^N \alpha}{\sqrt{2\pi}} \sin(\omega_0 x) \frac{\sin^N\left(\frac{\beta x}{2}\right)}{x^N},$$

with $\phi_{\mathbb{P}} = \phi_{\mathbb{Q}} + \theta = \phi_{\mathbb{Q}} + i\theta_I$ where $\theta_I = \text{Im}[\theta]$ and $\phi_{\mathbb{P}} \in L^1(\mathbb{R})$. So, we have constructed $\mathbb{P} \neq \mathbb{Q}$, such that $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$. Figure 3.1 shows the plots of ψ, $\widehat{\psi}$, θ, $\theta^\vee$, q, $\phi_{\mathbb{Q}}$, p and $|\phi_{\mathbb{P}}|$ for $\beta = 2\pi$, $N = 2$, $\omega_0 = 4\pi$ and $\alpha = \frac{1}{50}$.

We now prove Theorem 3.17.

[5] Since the sinc kernel is not in $L^1(\mathbb{R})$, its Fourier transform does not exist in the L^1-sense. However, its Fourier transform exists in the L^2-sense. See Section C.1.2 for details.



(a)

(a′)

(b)

(b′)

(c)

(c′)

(d)

(d′)

Figure 3.1: (a-a′) ψ and its Fourier spectrum $\widehat{\psi}$, (b-b′) $\theta^\vee$ and $i\theta$, (c-c′) the Cauchy distribution, q and its characteristic function $\phi_{\mathbb{Q}}$, and (d-d′) $p = q + \theta^\vee$ and $|\phi_{\mathbb{P}}|$. See Example 3.23 for details.

Proof of Theorem 3.17. Suppose $\exists\,\mathbb{P}\neq\mathbb{Q}$, $\mathbb{P},\mathbb{Q}\in\mathscr{P}_1(\mathbb{R}^d)$ such that $\gamma_k(\mathbb{P},\mathbb{Q})=0$. Since any positive Borel measure on $\mathbb{R}^d$ is a distribution [65, p. 157], $\mathbb{P}$ and $\mathbb{Q}$ can be treated as distributions with compact support. By the Paley-Wiener theorem (Theorem C.9), $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ are restrictions to $\mathbb{R}^d$ of entire functions on $\mathbb{C}^d$. Let $\theta:=\phi_{\mathbb{P}}-\phi_{\mathbb{Q}}$. Since $\gamma_k(\mathbb{P},\mathbb{Q})=0$, we have from (3.7) that $\int_{\mathbb{R}^d}|\theta(\omega)|^2\,d\Lambda(\omega)=0$. From Remark 3.22, it follows that $\text{supp}(\theta)\subset\text{cl}(\mathbb{R}^d\backslash\text{supp}(\Lambda))$. Since $\text{supp}(\Lambda)$ has a non-empty interior, we have $\text{supp}(\theta)\subsetneq\mathbb{R}^d$. Thus, there exists an open set, $U\subset\mathbb{R}^d$ such that $\theta(x)=0,\,\forall\,x\in U$. Since θ is analytic on $\mathbb{R}^d$, we have $\theta=0$, which means $\phi_{\mathbb{P}}=\phi_{\mathbb{Q}}\Rightarrow\mathbb{P}=\mathbb{Q}$, leading to a contradiction. So, there does not exist $\mathbb{P}\neq\mathbb{Q}$, $\mathbb{P},\mathbb{Q}\in\mathscr{P}_1(\mathbb{R}^d)$ such that $\gamma_k(\mathbb{P},\mathbb{Q})=0$, and k is therefore characteristic to $\mathscr{P}_1(\mathbb{R}^d)$. □

The condition that $\text{supp}(\Lambda)$ has a non-empty interior is important for Theorem 3.17 to hold. In the following, we provide a simple example to show that $\mathbb{P}\neq\mathbb{Q}$, $\mathbb{P},\mathbb{Q}\in\mathscr{P}_1(\mathbb{R}^d)$ can be constructed such that $\gamma_k(\mathbb{P},\mathbb{Q})=0$, if k is a periodic translation invariant kernel for which $\text{int}(\text{supp}(\Lambda))=\emptyset$.

Example 3.24. *Let $\mathbb{Q}$ be a uniform distribution on $[-\beta,\beta]\subset\mathbb{R}$, that is, $q(x)=\frac{1}{2\beta}\mathbb{1}_{[-\beta,\beta]}(x)$ with its characteristic function, $\phi_{\mathbb{Q}}(\omega)=\frac{1}{\beta\sqrt{2\pi}}\frac{\sin(\beta\omega)}{\omega}\in L^2(\mathbb{R})$. Let ψ be the Dirichlet kernel with period τ, where $\tau\le\beta$, that is, $\psi(x)=\frac{\sin\frac{(2l+1)\pi x}{\tau}}{\sin\frac{\pi x}{\tau}}$ and $\widehat{\psi}(\omega)=\sqrt{2\pi}\sum_{j=-l}^{l}\delta\left(\omega-\frac{2\pi j}{\tau}\right)$ with $\text{supp}(\widehat{\psi})=\left\{\frac{2\pi j}{\tau}\,:\,j\in\{0,\pm1,\ldots,\pm l\}\right\}$. Clearly, $\text{supp}(\widehat{\psi})$ has an empty interior. Let θ be*

$$\theta(\omega)=\frac{8\sqrt{2}\alpha}{i\sqrt{\pi}}\sin\left(\frac{\omega\tau}{2}\right)\frac{\sin^2\left(\frac{\omega\tau}{4}\right)}{\tau\omega^2},$$

with $\alpha\le\frac{1}{2\beta}$. It is easy to verify that $\theta\in L^1(\mathbb{R})\cap L^2(\mathbb{R})\cap C_b(\mathbb{R})$, so θ satisfies (i) in Lemma 3.21. Since $\theta(\omega)=0$ at $\omega=\frac{2\pi l}{\tau}$, $l\in\mathbb{Z}$, $\text{supp}(\theta)\cap\text{supp}(\widehat{\psi})\subset\text{supp}(\widehat{\psi})$ is a set of Lebesgue measure zero, so (iii) and (iv) in Lemma 3.21 are satisfied. $\theta^\vee$ is given by

$$\theta^\vee(x)=\begin{cases}\frac{2\alpha\left|x+\frac{\tau}{2}\right|}{\tau}-\alpha, & -\tau\le x\le0\\ \alpha-\frac{2\alpha\left|x-\frac{\tau}{2}\right|}{\tau}, & 0\le x\le\tau\\ 0, & \textit{otherwise,}\end{cases}$$

where $\theta^\vee \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ *satisfies (ii) in Lemma 3.21. Now, consider* $p = q + \theta^\vee$*, which is given as*

$$p(x) = \begin{cases} \frac{1}{2\beta}, & x \in [-\beta, -\tau] \cup [\tau, \beta] \\ \frac{2\alpha\left|x+\frac{\tau}{2}\right|}{\tau} + \frac{1}{2\beta} - \alpha, & x \in [-\tau, 0] \\ \alpha + \frac{1}{2\beta} - \frac{2\alpha\left|x-\frac{\tau}{2}\right|}{\tau}, & x \in [0, \tau] \\ 0, & \textit{otherwise.} \end{cases}$$

Clearly, $p(x) \geq 0, \forall\, x$ *and* $\int_{\mathbb{R}} p(x)\, dx = 1$. $\phi_{\mathbb{P}} = \phi_{\mathbb{Q}} + \theta = \phi_{\mathbb{Q}} + i\theta_I$ *where* $\theta_I = \mathrm{Im}[\theta]$ *and* $\phi_{\mathbb{P}} \in L^2(\mathbb{R})$. *We have therefore constructed* $\mathbb{P} \neq \mathbb{Q}$*, such that* $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$*, where* $\mathbb{P}$ *and* $\mathbb{Q}$ *are compactly supported in* $\mathbb{R}$ *with characteristic functions in* $L^2(\mathbb{R})$*, that is,* $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_1(\mathbb{R}^d)$. *Figure 3.2 shows the plots of* ψ, $\widehat{\psi}$, θ, $\theta^\vee$, q, $\phi_{\mathbb{Q}}$, p *and* $|\phi_{\mathbb{P}}|$ *for* $\tau = 2$, $l = 2$, $\beta = 3$ *and* $\alpha = \frac{1}{8}$.

We now present the proof of Theorem 3.19, which is similar to that of Theorem 3.13.

Proof of Theorem 3.19. ($\Leftarrow$) From (3.6), we have

$$\begin{aligned} \gamma_k^2(\mathbb{P}, \mathbb{Q}) &= \iint_{\mathbb{T}^d} \psi(x-y)\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y) \\ &\overset{(a)}{=} \iint_{\mathbb{T}^d} \sum_{n \in \mathbb{Z}^d} A_\psi(n)\, e^{i\langle x-y, n\rangle}\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y) \\ &\overset{(b)}{=} \sum_{n \in \mathbb{Z}^d} A_\psi(n) \left| \int_{\mathbb{T}^d} e^{-i\langle x, n\rangle}\, d(\mathbb{P}-\mathbb{Q})(x) \right|^2 \\ &\overset{(c)}{=} (2\pi)^{2d} \sum_{n \in \mathbb{Z}^d} A_\psi(n)\, |A_{\mathbb{P}}(n) - A_{\mathbb{Q}}(n)|^2, \end{aligned} \tag{3.13}$$

where we have invoked (2.4) in (a), Fubini's theorem in (b) and

$$A_{\mathbb{P}}(n) := \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} e^{-i\langle n, x\rangle}\, d\mathbb{P}(x),\ n \in \mathbb{Z}^d,$$

in (c). $A_{\mathbb{P}}$ is the Fourier transform of $\mathbb{P}$ in $\mathbb{T}^d$. Since $A_\psi(0) \geq 0$ and $A_\psi(n) > 0, \forall\, n \neq 0$, we have $A_{\mathbb{P}}(n) = A_{\mathbb{Q}}(n), \forall\, n$. Therefore, by the uniqueness theorem of Fourier transform, we have $\mathbb{P} = \mathbb{Q}$.

($\Rightarrow$) Proving the necessity is equivalent to proving that if $A_\psi(0) \geq 0$, $A_\psi(n) >$

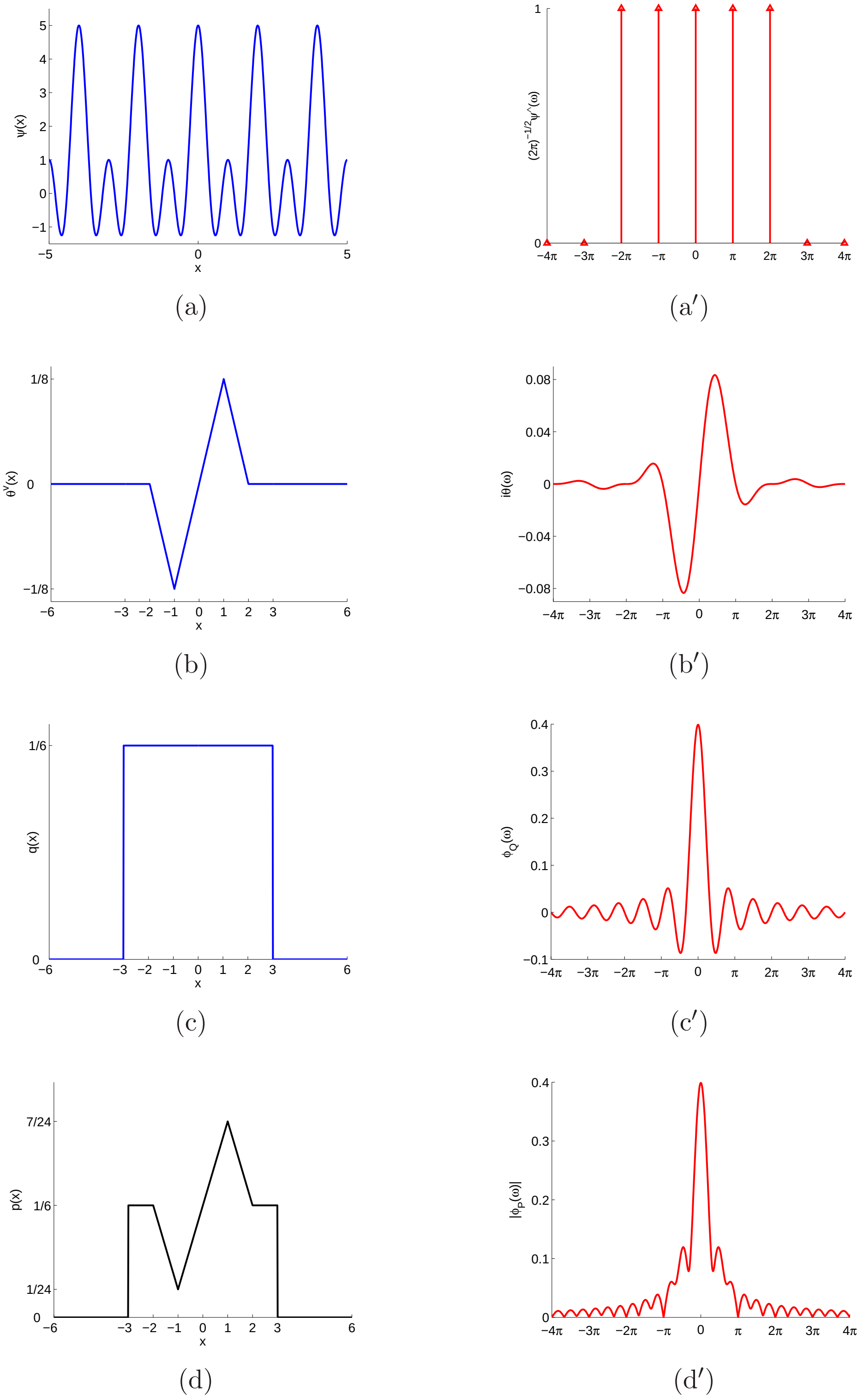


Figure 3.2: (a-a′) ψ and its Fourier spectrum $\widehat{\psi}$, (b-b′) $\theta^{\vee}$ and $i\theta$, (c-c′) the uniform distribution, q and its characteristic function $\phi_{\mathbb{Q}}$, and (d-d′) $p = q + \theta^{\vee}$ and $|\phi_{\mathbb{P}}|$. See Example 3.24 for details.

$0, \forall\, n \neq 0$ is violated, then k is not characteristic, which is equivalent to showing that $\exists\, \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. Let $\mathbb{Q}$ be a uniform probability measure with $q(x) = \frac{1}{(2\pi)^d}, \forall\, x \in \mathbb{T}^d$. Let k be such that $A_\psi(n) = 0$ for some $n = n_0 \neq 0$. Define

$$A_{\mathbb{P}}(n) := \begin{cases} A_{\mathbb{Q}}(n), & n \neq \pm n_0 \\ A_{\mathbb{Q}}(n) + \theta(n), & n = \pm n_0 \end{cases},$$

where $A_{\mathbb{Q}}(n) = \frac{1}{(2\pi)^d}\delta_{0n}$ and $\theta(-n_0) = \overline{\theta(n_0)}$. So,

$$p(x) = \sum_{n \in \mathbb{Z}^d} A_{\mathbb{P}}(n) e^{i\langle x, n\rangle} = \frac{1}{(2\pi)^d} + \theta(n_0) e^{i\langle x, n_0\rangle} + \theta(-n_0) e^{-i\langle x, n_0\rangle}.$$

Choose $\theta(n_0) = i\alpha$, $\alpha \in \mathbb{R}$. Then, $p(x) = \frac{1}{(2\pi)^d} - 2\alpha \sin(\langle x, n_0\rangle)$. It is easy to check that p integrates to one. Choosing $|\alpha| \leq \frac{1}{2(2\pi)^d}$ ensures that $p(x) \geq 0, \forall\, x \in \mathbb{T}^d$. By using $A_{\mathbb{P}}(n)$ in (3.13), it is clear that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. Therefore, $\exists\, \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$, which means k is not characteristic. □

3.3 Dissimilar Distributions with Small MMD

So far, we have studied different characterizations for the kernel k such that γ_k is a metric on $M_+^1(\mathcal{X})$. As mentioned in Chapter 1, the metric property of γ_k is crucial in many statistical inference applications like hypothesis testing. Therefore, in practice, it is important to use characteristic kernels. However, in this section, we show that characteristic kernels, while guaranteeing γ_k to be a metric on $M_+^1(\mathcal{X})$, may nonetheless have difficulty in distinguishing certain distributions on the basis of finite samples. More specifically, in Theorem 3.27 we show that for a given kernel k and for any $\varepsilon > 0$, there exist $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) < \varepsilon$. Before proving the result, we motivate it through the following example.

Example 3.25. *Let $\mathbb{P}$ be absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}$ with the Radon-Nikodym derivative defined as*

$$p(x) = q(x) + \alpha q(x) \sin(\nu\pi x), \tag{3.14}$$

where q is the Radon-Nikodym derivative of $\mathbb{Q}$ w.r.t. the Lebesgue measure satisfying $q(x) = q(-x), \forall\, x$ and $\alpha \in [-1, 1]\backslash\{0\}$, $\nu \in \mathbb{R}\backslash\{0\}$. It is obvious that $\mathbb{P} \neq \mathbb{Q}$.

The characteristic function of $\mathbb{P}$ *is given as*

$$\phi_{\mathbb{P}}(\omega) = \phi_{\mathbb{Q}}(\omega) - \frac{i\alpha}{2}\left[\phi_{\mathbb{Q}}(\omega - \nu\pi) - \phi_{\mathbb{Q}}(\omega + \nu\pi)\right], \, \omega \in \mathbb{R},$$

where $\phi_{\mathbb{Q}}$ *is the characteristic function associated with* $\mathbb{Q}$*. Note that with increasing* $|\nu|$, p *has higher frequency components in its Fourier spectrum, as shown in Figure 3.3. In Figure 3.3, (a-c) show the plots of* p *when* $q = U[-1, 1]$ *(uniform distribution) and (a′-c′) show the plots of* p *when* $q = N(0, 2)$ *(zero mean normal distribution with variance* 2*) for* $\nu = 0, 2$ *and* 7.5 *with* $\alpha = \frac{1}{2}$.

Consider the B_1*-spline kernel on* $\mathbb{R}$ *given by* $k(x, y) = \psi(x - y)$ *where*

$$\psi(x) = \begin{cases} 1 - |x|, & |x| \le 1 \\ 0, & otherwise \end{cases}, \tag{3.15}$$

with its Fourier transform given by

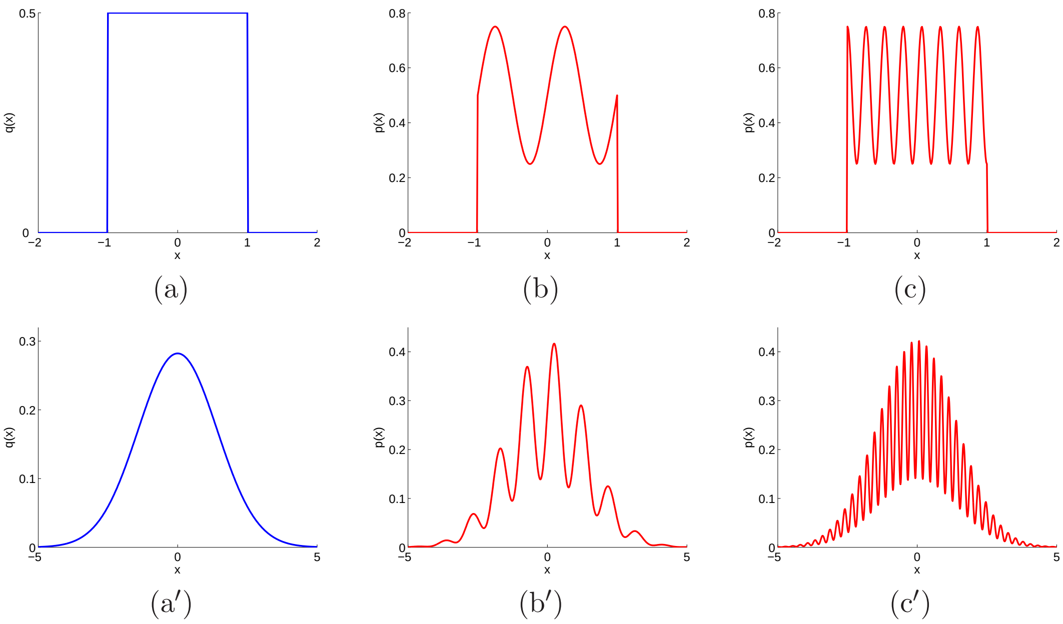
$$\widehat{\psi}(\omega) = \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{\sin^2 \frac{\omega}{2}}{\omega^2}.$$

Since ψ *is characteristic to* $M^1_+(\mathbb{R})$, $\gamma_k(\mathbb{P}, \mathbb{Q}) > 0$ *(see Theorem 3.13). However, it would be of interest to study the behavior of* $\gamma_k(\mathbb{P}, \mathbb{Q})$ *as a function of* ν*. We study the behavior of* $\gamma^2_k(\mathbb{P}, \mathbb{Q})$ *through its unbiased, consistent estimator,*[6] $\gamma^2_{k,u}(m, m)$ *as considered by [37, Lemma 7].*

Figure 3.4(a) shows the behavior of $\gamma^2_{k,u}(m, m)$ *as a function of* ν *for* $q = U[-1, 1]$ *and* $q = N(0, 2)$ *using the* B_1*-spline kernel in (3.15). Since the Gaussian kernel,* $k(x, y) = e^{-(x-y)^2}$ *is also a characteristic kernel, its effect on the behavior of* $\gamma^2_{k,u}(m, m)$ *is shown in Figure 3.4(b) in comparison to that of the* B_1*-spline kernel.*

In Figure 3.4, we observe two circumstances under which γ^2_k *may be small. First,* $\gamma^2_{k,u}(m, m)$ *decays with increasing* $|\nu|$*, and can be made as small as desired by choosing a sufficiently large* $|\nu|$*. Second, in Figure 3.4(a),* $\gamma^2_{k,u}(m, m)$ *has troughs at* $\nu = \frac{\omega_0}{\pi}$ *where* $\omega_0 = \{\omega : \widehat{\psi}(\omega) = 0\}$*. Since* $\gamma^2_{k,u}(m, m)$ *is a consistent estimate*

[6] Let $\{X_j\}_{j=1}^m$ and $\{Y_j\}_{j=1}^m$ be random samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$ respectively. An unbiased *empirical estimate* of $\gamma^2_k(\mathbb{P}, \mathbb{Q})$, denoted as $\gamma^2_{k,u}(m, m)$ is given by $\gamma^2_{k,u}(m, m) = \frac{1}{m(m-1)} \sum_{l \neq j}^m h(Z_l, Z_j)$, which is a one-sample U-statistic [69, Chapter 5] with $h(Z_l, Z_j) := k(X_l, X_j) + k(Y_l, Y_j) - k(X_l, Y_j) - k(X_j, Y_l)$, where $Z_1, \ldots, Z_m$ are m i.i.d. random variables with $Z_j := (X_j, Y_j)$. See Chapter 5 and [37, Lemma 7] for more details on the estimation of γ_k from finite samples.



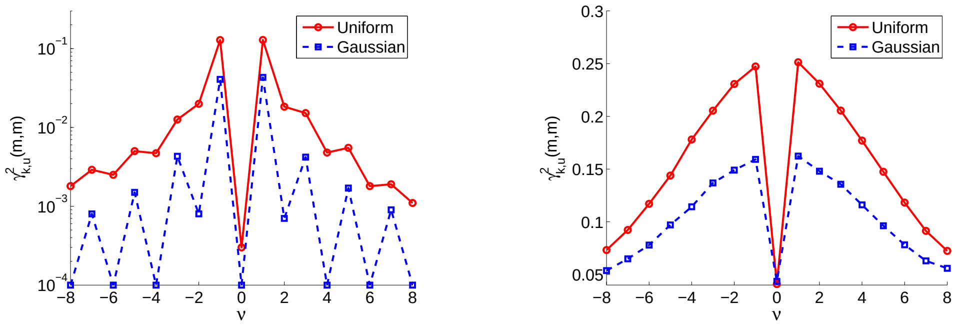
(a) (b) (c)

(a′) (b′) (c′)

Figure 3.3: (a) $q = U[-1,1]$, (a′) $q = N(0,2)$. (b-c) and (b′-c′) denote $p(x)$ computed as $p(x) = q(x) + \frac{1}{2}q(x)\sin(\nu\pi x)$ with $q = U[-1,1]$ and $q = N(0,2)$ respectively. ν is chosen to be 2 in (b,b′) and 7.5 in (c,c′). See Example 3.25 for details.

of $\gamma_k^2(\mathbb{P},\mathbb{Q})$, one would expect similar behavior from $\gamma_k^2(\mathbb{P},\mathbb{Q})$. This means that, although the B_1-spline kernel is characteristic to $M_+^1(\mathbb{R})$, in practice, it becomes harder to distinguish between $\mathbb{P}$ and $\mathbb{Q}$ with finite samples, when $\mathbb{P}$ is constructed as in (3.14) with $\nu = \frac{\omega_0}{\pi}$. In fact, one can observe from a straightforward spectral argument that the troughs in $\gamma_k^2(\mathbb{P},\mathbb{Q})$ can be made arbitrarily deep by widening q, when q is Gaussian.

For characteristic kernels, although $\gamma_k(\mathbb{P},\mathbb{Q}) > 0$ when $\mathbb{P} \neq \mathbb{Q}$, Example 3.25 demonstrates that one can construct distributions such that $\gamma_{k,u}^2(m,m)$ is indistinguishable from zero with high probability, for a given sample size m. Below, in Theorem 3.27, we explicitly construct $\mathbb{P} \neq \mathbb{Q}$ such that $|\mathbb{P}\varphi_l - \mathbb{Q}\varphi_l|$ is large for some large l, but $\gamma_k(\mathbb{P},\mathbb{Q})$ is arbitrarily small, making it hard to detect a non-zero value of $\gamma_k(\mathbb{P},\mathbb{Q})$ based on finite samples. Here, $\varphi_l \in L^2(\mathcal{X})$ represents the bounded orthonormal eigenfunctions of a positive definite integral operator associated with k and $\mathbb{P}\varphi_l := \int_{\mathcal{X}} \varphi_l \, d\mathbb{P}$. Based on this theorem, for example, in Example 3.25, the decay mode of γ_k for large $|\nu|$ can be investigated.



(a)

(b)

Figure 3.4: Behavior of the empirical estimate of $\gamma_k^2(\mathbb{P},\mathbb{Q})$ w.r.t. ν for (a) the B_1-spline kernel and (b) the Gaussian kernel. $\mathbb{P}$ is constructed from $\mathbb{Q}$ as defined in (3.14). "Uniform" corresponds to $\mathbb{Q} = U[-1,1]$ and "Gaussian" corresponds to $\mathbb{Q} = N(0,2)$. $m = 1000$ samples are generated from $\mathbb{P}$ and $\mathbb{Q}$ to estimate $\gamma_k^2(\mathbb{P},\mathbb{Q})$ through $\gamma^2_{k,u}(m,m)$. This is repeated 100 times and the average $\gamma^2_{k,u}(m,m)$ is plotted in both figures. Since the quantity of interest is the average behavior of $\gamma^2_{k,u}(m,m)$, we omit the error bars. See Example 3.25 for details.

The construction of $\mathbb{P}$ for a given $\mathbb{Q}$ such that $\gamma_k(\mathbb{P},\mathbb{Q})$ is small, though not zero, can be intuitively understood by re-writing the result of Proposition 3.2 as

$$\gamma_k(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}_k}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{H}} \frac{|\mathbb{P}f-\mathbb{Q}f|}{\|f\|_{\mathcal{H}}},$$

where $\mathbb{P}f := \int_{\mathcal{X}} f\,d\mathbb{P}$. When $\mathbb{P}\neq\mathbb{Q}$, $|\mathbb{P}f-\mathbb{Q}f|$ can be large for some $f\in\mathcal{H}$. However, $\gamma_k(\mathbb{P},\mathbb{Q})$ can be made small by selecting $\mathbb{P}$ such that the maximization of $\frac{|\mathbb{P}f-\mathbb{Q}f|}{\|f\|_{\mathcal{H}}}$ over $\mathcal{H}$ requires an f with large $\|f\|_{\mathcal{H}}$. More specifically, higher order eigenfunctions of the kernel (φ_l for large l) have large RKHS norms, so, if they are prominent in $\mathbb{P}$ and $\mathbb{Q}$ (i.e., highly non-smooth distributions), one can expect $\gamma_k(\mathbb{P},\mathbb{Q})$ to be small even when there exists an l for which $|\mathbb{P}\varphi_l-\mathbb{Q}\varphi_l|$ is large. To this end, we need the following lemma, which we quote from [39, Lemma 4].

Lemma 3.26 ([39]). *Let $\mathcal{F}$ be the unit ball in an RKHS $(\mathcal{H},k)$ defined on a compact topological space, $\mathcal{X}$, with k being measurable. Let $\varphi_l\in L^2(\mathcal{X},\mu)$ be absolutely bounded orthonormal eigenfunctions and λ_l be the corresponding eigenvalues (arranged in decreasing order for increasing l) of a positive definite integral operator associated with k and a σ-finite measure, μ. Assume λ_l^{-1} increases super-linearly with l. Then, for $f\in\mathcal{F}$ where $f(x)=\sum_{j=1}^\infty \widetilde{f}_j\varphi_j(x)$, $\widetilde{f}_j := \langle f,\varphi_j\rangle_{L^2(\mathcal{X},\mu)}$, we have $\sum_{j=1}^\infty|\widetilde{f}_j|<\infty$ and for every $\varepsilon>0$, $\exists\, l_0\in\mathbb{N}$ such that $|\widetilde{f}_l|<\varepsilon$ if $l>l_0$.*

Theorem 3.27 ($\mathbb{P} \neq \mathbb{Q}$ can have arbitrarily small γ_k)**.** *Suppose the conditions in Lemma 3.26 hold. Then there exist probability measures $\mathbb{P} \neq \mathbb{Q}$ defined on $\mathcal{X}$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) < \varepsilon$ for any arbitrarily small $\varepsilon > 0$.*

Proof. Suppose q be the Radon-Nikodym derivative associated with $\mathbb{Q}$ w.r.t. the σ-finite measure, μ (see Lemma 3.26). Let us construct $p(x) = q(x) + \alpha_l e(x) + \tau\varphi_l(x)$ where $e(x) = \mathbb{1}_{\mathcal{X}}(x)$. For $\mathbb{P}$ to be a probability measure, the following conditions need to be satisfied:

$$\begin{aligned} &\int_{\mathcal{X}} [\alpha_l e(x) + \tau\varphi_l(x)]\, d\mu(x) = 0, \\ &\min_{x\in\mathcal{X}} [q(x) + \alpha_l e(x) + \tau\varphi_l(x)] \geq 0. \end{aligned} \tag{3.16}$$

Expanding $e(x)$ and $f(x)$ in the orthonormal basis $\{\varphi_l\}_{l=1}^{\infty}$, we get

$$e(x) = \sum_{l=1}^{\infty} \widetilde{e}_l \varphi_l(x) \text{ and } f(x) = \sum_{l=1}^{\infty} \widetilde{f}_l \varphi_l(x),$$

where $\widetilde{e}_l := \langle e, \varphi_l\rangle_{L^2(\mathcal{X},\mu)}$ and $\widetilde{f}_l := \langle f, \varphi_l\rangle_{L^2(\mathcal{X},\mu)}$. Therefore,

$$\begin{aligned} \mathbb{P}f - \mathbb{Q}f &= \int_{\mathcal{X}} f(x)\,[\alpha_l e(x) + \tau\varphi_l(x)]\, d\mu(x) \\ &= \int_{\mathcal{X}} \left[\alpha_l \sum_{j=1}^{\infty} \widetilde{e}_j \varphi_j(x) + \tau\varphi_l(x)\right] \left[\sum_{t=1}^{\infty} \widetilde{f}_t \varphi_t(x)\right] d\mu(x) \\ &= \alpha_l \sum_{j=1}^{\infty} \widetilde{e}_j \widetilde{f}_j + \tau \widetilde{f}_l, \end{aligned} \tag{3.17}$$

where we used the fact that $\langle \varphi_j, \varphi_t\rangle_{L^2(\mathcal{X},\mu)} = \delta_{jt}$ (here, δ is used in the Kronecker sense). Rewriting (3.16) and substituting for $e(x)$ gives

$$\int_{\mathcal{X}} [\alpha_l e(x) + \tau\varphi_l(x)]\, d\mu(x) = \int_{\mathcal{X}} e(x)[\alpha_l e(x) + \tau\varphi_l(x)]\, d\mu(x) = \alpha_l \sum_{j=1}^{\infty} \widetilde{e}_j^2 + \tau\widetilde{e}_l = 0,$$

which implies

$$\alpha_l = -\frac{\tau\widetilde{e}_l}{\sum_{j=1}^{\infty} \widetilde{e}_j^2}. \tag{3.18}$$

Now, let us consider $\mathbb{P}\varphi_t - \mathbb{Q}\varphi_t = \alpha_l \widetilde{e}_t + \tau\delta_{tl}$. Substituting for α_l gives

$$\mathbb{P}\varphi_t - \mathbb{Q}\varphi_t = \tau\delta_{tl} - \tau\frac{\widetilde{e}_t\widetilde{e}_l}{\sum_{j=1}^{\infty} \widetilde{e}_j^2} = \tau\delta_{tl} - \tau\rho_{tl},$$

where $\rho_{tl} := \frac{\widetilde{e}_t \widetilde{e}_l}{\sum_{j=1}^{\infty} \widetilde{e}_j^2}$. By Lemma 3.26, $\sum_{l=1}^{\infty} |\widetilde{e}_l| < \infty \Rightarrow \sum_{j=1}^{\infty} \widetilde{e}_j^2 < \infty$, and choosing large enough l gives $|\rho_{tl}| < \eta$, $\forall\, t$, for any arbitrary $\eta > 0$. Therefore, $|\mathbb{P}\varphi_t - \mathbb{Q}\varphi_t| > \tau - \eta$ for $t = l$ and $|\mathbb{P}\varphi_t - \mathbb{Q}\varphi_t| < \eta$ for $t \neq l$, which means $\mathbb{P} \neq \mathbb{Q}$. In the following, we prove that $\gamma_k(\mathbb{P}, \mathbb{Q})$ can be arbitrarily small, though non-zero.

Recall that $\gamma_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\mathbb{P}f - \mathbb{Q}f|$. Substituting (3.18) in (3.17) and replacing $|\mathbb{P}f - \mathbb{Q}f|$ by (3.17) in $\gamma_k(\mathbb{P}, \mathbb{Q})$, we have

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \sup_{\{\widetilde{f}_j\}_{j=1}^{\infty}} \left\{ \tau \sum_{j=1}^{\infty} \nu_{jl} \widetilde{f}_j \ : \ \sum_{j=1}^{\infty} \frac{\widetilde{f}_j^2}{\lambda_j} \leq 1 \right\}, \tag{3.19}$$

where we used the definition of RKHS norm as $\|f\|_{\mathcal{H}}^2 := \sum_{j=1}^{\infty} \frac{\widetilde{f}_j^2}{\lambda_j}$ [81, Theorem 4.51] and $\nu_{jl} := \delta_{jl} - \rho_{jl}$. (3.19) is a convex quadratically constrained quadratic program in $\{\widetilde{f}_j\}_{j=1}^{\infty}$. Solving the Lagrangian yields $\widetilde{f}_j = \frac{\nu_{jl}\lambda_j}{\sqrt{\sum_{j=1}^{\infty} \nu_{jl}^2 \lambda_j}}$. Therefore,

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \tau \sqrt{\sum_{j=1}^{\infty} \nu_{jl}^2 \lambda_j} = \tau \sqrt{\lambda_l - 2\rho_{ll}\lambda_l + \sum_{j=1}^{\infty} \rho_{jl}^2 \lambda_j} \xrightarrow{l \to \infty} 0,$$

because *(i)* by choosing sufficiently large l, $|\rho_{jl}| < \varepsilon$, $\forall\, j$, for any arbitrary $\varepsilon > 0$, and *(ii)* $\lambda_l \to 0$ as $l \to \infty$ [68, Theorem 2.10]. Therefore, we have constructed $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) < \varepsilon$ for any arbitrarily small $\varepsilon > 0$. □

3.4 Discussion

In this chapter, we have presented various interpretations of γ_k and studied the conditions (on k) under which it is a metric on $M_+^1(\mathcal{X})$. We showed that apart from universal kernels (in the sense of Steinwart [80]), a large family of bounded continuous kernels induces a metric, γ_k on $M_+^1(\mathcal{X})$: (a) integrally strictly pd kernels and (b) translation invariant kernels on $\mathbb{R}^d$ and $\mathbb{T}^d$ that have the support of their Fourier transform to be $\mathbb{R}^d$ and $\mathbb{Z}^d$ respectively. We also showed that there exist distinct distributions which will be considered close according to γ_k (whether or not the kernel is characteristic), and thus may be hard to distinguish based on finite samples.

We now discuss how kernels on $M^1_+(\mathcal{X})$ can be obtained from γ_k. As noted by Gretton et al. [37, Section 4], and following [41], γ_k is a *Hilbertian metric*[7] [8, Section 3.3] on $M^1_+(\mathcal{X})$: the associated kernel can be easily computed as[8]

$$\widetilde{K}(\mathbb{P},\mathbb{Q}) = \left\langle \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x), \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{Q}(x) \right\rangle_{\mathcal{H}} = \iint_{\mathcal{X}} k(x,y)\,d\mathbb{P}(x)\,d\mathbb{Q}(y),$$

where the pd kernel $\widetilde{K} : M^1_+(\mathcal{X}) \times M^1_+(\mathcal{X}) \to \mathbb{R}$ is a dot-product kernel on $M^1_+(\mathcal{X})$. Using the results in [8, Chapter 3, Theorems 2.2 and 2.3], Gaussian and inverse multi-quadratic kernels on $M^1_+(\mathcal{X})$ can be defined as

$$\widetilde{K}(\mathbb{P},\mathbb{Q}) = \exp\left(-\sigma\gamma_k^2(\mathbb{P},\mathbb{Q})\right),\ \sigma > 0 \ \text{ and } \ \widetilde{K}(\mathbb{P},\mathbb{Q}) = \left(\sigma + \gamma_k^2(\mathbb{P},\mathbb{Q})\right)^{-1},\ \sigma > 0$$

respectively. Further work on Hilbertian metrics and positive definite kernels on probability measures has been carried out by [40] and [27].

Bibliographic Notes

This chapter is based on joint work with Kenji Fukumizu, Arthur Gretton, Gert Lanckriet and Bernhard Schölkopf, which appeared in [75, 78, 79]. The dissertation author was the primary investigator and author of these papers.

[7] A metric ρ on $\mathcal{X}$ is said to be Hilbertian if there exists a Hilbert space, $\mathcal{H}$ and a mapping Φ such that $\rho(x,y) = \|\Phi(x) - \Phi(y)\|_{\mathcal{H}},\ \forall\, x,y \in \mathcal{X}$.

[8] Based on Section 3.3 (p. 81) and Lemma 3.2.1 of [8], it can be stated that if $(\mathcal{X},\rho)$ is a non-empty set, $x_0 \in \mathcal{X}$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric kernel defined as $k(x,y) = \frac{1}{2}\left(\rho^2(x,x_0) + \rho^2(y,x_0) - \rho^2(x,y) - \rho^2(x_0,x_0)\right)$, then k is pd if and only if ρ is Hilbertian.

Table 3.1: The table should be read as: If “Property” is satisfied on $\mathcal{X}$, then k is characteristic (or not) to $\mathscr{Q}(\mathcal{X})$. $M_+^1(\mathcal{X})$ is the set of Borel probability measures on a topological space, $\mathcal{X}$. When $\mathcal{X}=\mathbb{R}^d$, $k(x,y)=\psi(x-y)$, where ψ is a bounded, continuous pd function on $\mathbb{R}^d$. ψ is the Fourier transform of a finite nonnegative Borel measure, Λ, and $\Omega := \text{supp}(\Lambda)$. $\mathscr{P}_1(\mathbb{R}^d) := \{\mathbb{P}\in M_+^1(\mathbb{R}^d) : \phi_{\mathbb{P}} \in L^1(\mathbb{R}^d)\cup L^2(\mathbb{R}^d),\, \mathbb{P}\ll\lambda \text{ and } \text{supp}(\mathbb{P}) \text{ is compact}\}$, where $\phi_{\mathbb{P}}$ is the characteristic function of $\mathbb{P}$ and λ is the Lebesgue measure. $\mathbb{P}\ll\lambda$ denotes that $\mathbb{P}$ is absolutely continuous w.r.t. λ. For a radial kernel k on $\mathbb{R}^d$, i.e., $k(x,y)=\int_{[0,\infty)} e^{-t\|x-y\|_2^2}\,d\nu(t)$, ν is a finite nonnegative Borel measure on $[0,\infty)$. When $\mathcal{X}=\mathbb{T}^d$, $k(x,y)=\psi(x-y)$, where ψ is a bounded, continuous pd function on $\mathbb{T}^d$. $\{A_\psi(n)\}_{n\in\mathbb{Z}^d}$ are the Fourier series coefficients of ψ which are nonnegative and summable (see (2.4) for details).

Summary of Main Results

Property	$\mathscr{Q}(\mathcal{X})$	Characteristic	Reference
k is integrally strictly pd	$M_+^1(\mathcal{X})$	Yes	Theorem 3.10
$\Omega=\mathbb{R}^d$	$M_+^1(\mathbb{R}^d)$	Yes	Theorem 3.13
supp(ψ) is compact	$M_+^1(\mathbb{R}^d)$	Yes	Corollary 3.14
$\Omega\subsetneq\mathbb{R}^d$, int$(\Omega)\neq\emptyset$	$\mathscr{P}_1(\mathbb{R}^d)$	Yes	Theorem 3.17
$\Omega\subsetneq\mathbb{R}^d$	$M_+^1(\mathbb{R}^d)$	No	Theorem 3.13
supp$(\nu)\neq\{0\}$	$M_+^1(\mathbb{R}^d)$	Yes	Corollary 3.16
supp$(\nu)=\{0\}$	$M_+^1(\mathbb{R}^d)$	No	Corollary 3.16
$A_\psi(0)\geq 0$, $A_\psi(n)>0$, $\forall\, n\neq 0$	$M_+^1(\mathbb{T}^d)$	Yes	Theorem 3.19
$\exists\, n\neq 0\,\|\, A_\psi(n)=0$	$M_+^1(\mathbb{T}^d)$	No	Theorem 3.19

4 Universality, Characteristic Kernels and Other Notions

In Chapter 3, we have presented various characterizations of characteristic kernels, which are easily checkable compared with characterizations proposed in earlier literature [30, 31, 37]. Using all these characterizations, in this chapter, we discuss and summarize the relation of characteristic kernels to other notions of pd kernels like *universal*, strictly pd, integrally strictly pd and conditionally strictly pd. Throughout this chapter, we assume $\mathcal{X}$ to be a Polish space,[9] the reason for which is discussed in footnote 12.

The chapter is organized as follows. In Section 4.1, we discuss in detail the notion of *universality*, which is shown to be related to the RKHS embedding of finite signed Borel measures. The relation between universal and characteristic kernels is discussed in Section 4.2, while both these notions are related to strictly pd, integrally strictly pd and conditionally strictly pd kernels in Section 4.3. This is done by first reviewing all the existing characterizations for universal and characteristic kernels, which is then used to study not only the relation between them but also their relation to other notions of pd kernels. Since the existing characterizations do not explain the complete relationship between all these various notions of pd kernels, we raise questions at the end of each section that need to be addressed to obtain a complete understanding of the relationships between all these notions, which are then addressed in Section 4.4 by deriving new results. A sum-

[9] A topological space $(\mathcal{X}, \tau)$ is called a Polish space if the topology τ has a countable basis and there exists a complete metric defining τ. An example of a Polish space is $\mathbb{R}^d$ endowed with its usual topology.

mary of the relation between all these notions of pd kernels is shown in Figure 4.1. For example, all these notions of pd kernels are shown to be equivalent for many popular kernels like Gaussian, Laplacian, inverse multi-quadratic, etc.

4.1 Universal Kernels

As mentioned in Chapter 1, RKHS based learning paradigm is broadly established as an easy way to construct nonlinear algorithms from linear ones, by embedding data points into an RKHS [68, 70]. In this approach to learning, the kernel-based algorithms (for classification/regression) generally invoke the *representer theorem* [44, 67] and learn a function (in an RKHS) that has the representation,

$$f := \sum_{j\in\mathbb{N}_n} c_j k(\cdot, x_j), \tag{4.1}$$

where $\mathbb{N}_n := \{1, 2, \dots, n\}$, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric pd kernel on $\mathcal{X}$ and $\{c_j : j \in \mathbb{N}_n\} \subset \mathbb{R}$ are parameters typically obtained from training data, $\{x_j : j \in \mathbb{N}_n\} \subset \mathcal{X}$. As noted in [54], one can ask whether the function, f in (4.1) approximates any real-valued target function arbitrarily *well* as the number of summands increases without bound. This is an important question to consider because if the answer is affirmative, then the kernel-based learning algorithm can be *consistent* in the sense that for any target function, $f^\star$, the discrepancy between f (which is learned from the training data) and $f^\star$ goes to zero (in some appropriate sense) as the sample size goes to infinity. Since the linear hull of $\{k(\cdot, x) : x \in \mathcal{X}\}$ is dense in the RKHS, $\mathcal{H}$ associated with k [4], and assuming that the kernel-based algorithm makes f "converge to an appropriate function" in $\mathcal{H}$ as $n \to \infty$, the above question of approximating $f^\star$ arbitrarily *well* by f in (4.1) as n goes to infinity is equivalent to the question of whether $\mathcal{H}$ is rich enough to approximate any $f^\star$ arbitrarily *well* (such an RKHS is referred to as a universal RKHS and the corresponding kernel as a universal kernel). Depending on the choice of $\mathcal{X}$, the choice of target function space and the type of approximation, various notions of universality—c-universality [80], cc-universality [12, 54], c_0-universality [13, 77] and L_p-universality [13, 81]—have been proposed and characterized in literature. In the following sections, we define each of these notions of universality, review their existing characterizations and summarize the relation between them.

4.1.1 c-universality

[80] proposed the notion of c-universality, which is defined as follows:

Definition 4.1 (c-universality)**.** *A continuous pd kernel k on a compact Hausdorff space $\mathcal{X}$ is called c-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $C(\mathcal{X})$ w.r.t. the supremum norm, i.e., for every function $g \in C(\mathcal{X})$ and all $\epsilon > 0$, there exists an $f \in \mathcal{H}$ such that $\|f - g\|_\infty \leq \epsilon$.*

By applying the Stone-Weierstraß theorem [26, Theorem 4.45], Steinwart [80, Theorem 9] provided sufficient conditions for a kernel to be c-universal—a continuous kernel, k on a compact metric space, $\mathcal{X}$ is c-universal if the following hold: (a) $k(x, x) > 0, \forall\, x \in \mathcal{X}$, (b) there exists an injective feature map $\Phi : \mathcal{X} \to \ell_2$ of k with $\Phi(\mathcal{X}) = \{\Phi_n(\mathcal{X})\}_{n\in\mathbb{N}}$ and (c) $\text{span}\{\Phi_n : n \in \mathbb{N}\}$ is an algebra—using which the Gaussian kernel is shown to be c-universal on every compact subset of $\mathbb{R}^d$. Micchelli et al. [54, Proposition 1] related c-universality to the injective RKHS embedding of finite signed Borel measures by showing that k is c-universal if and only if

$$\mu \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x),\ \mu \in M_b(\mathcal{X}), \tag{4.2}$$

is injective.

4.1.2 cc-universality

One limitation in the notion of universality considered by Steinwart [80] is that $\mathcal{X}$ is assumed to be compact, which excludes many interesting spaces, such as $\mathbb{R}^d$ and infinite discrete sets. To overcome this limitation, Carmeli et al. [13, Definition 2, Theorem 3] and Sriperumbudur et al. [77] introduced the following notion of cc-universality which can handle non-compact Hausdorff spaces, $\mathcal{X}$.

Definition 4.2 (cc-universality)**.** *A continuous pd kernel k on a Hausdorff space $\mathcal{X}$ is said to be cc-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $C(\mathcal{X})$ endowed* *with the topology of compact convergence, i.e., for any compact set $\mathcal{Z} \subset \mathcal{X}$, for any $g \in C(\mathcal{Z})$ and all $\epsilon > 0$, there exists an $f \in \mathcal{H}|_{\mathcal{Z}}$ such that $\|f - g\|_\infty \le \epsilon$, where $\mathcal{H}|_{\mathcal{Z}} := \{f|_{\mathcal{Z}} : f \in \mathcal{H}\}$ is the restriction of $\mathcal{H}$ to $\mathcal{Z}$ and $f|_{\mathcal{Z}}$ is the restriction of f to $\mathcal{Z}$.*

Carmeli et al. [13, Theorem 3, Proposition 3 and Theorem 4] showed that a bounded continuous pd kernel, k is *cc*-universal if and only if the following embedding is injective for all $\mu \in M_{bc}(\mathcal{X})$ and some $p \in [1, \infty)$:

$$f \mapsto \int_{\mathcal{X}} k(\cdot, x) f(x)\, d\mu(x),\ f \in L^p(\mathcal{X}, \mu). \tag{4.3}$$

In addition, they [13, Remark 1] showed that k being *cc*-universal is equivalent to it being universal in the sense of [54] and [12]: for any compact $\mathcal{Z} \subset \mathcal{X}$, the set $\widetilde{K}(\mathcal{Z}) := \mathrm{cl}(\mathrm{span}\{k(\cdot, y) : y \in \mathcal{Z}\})$ is dense in $C(\mathcal{Z})$ in the supremum norm, which is shown by Micchelli et al. [54, Proposition 1] to be equivalent to the following embedding being injective:

$$\mu \mapsto \int_{\mathcal{Z}} k(\cdot, x)\, d\mu(x),\ \mu \in M_b(\mathcal{Z}). \tag{4.4}$$

Since (4.4) holds for any compact $\mathcal{Z} \subset \mathcal{X}$, the universality in the sense of [54] and [12] is equivalent to the following embedding being injective:

$$\mu \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x),\ \mu \in M_{bc}(\mathcal{X}), \tag{4.5}$$

where $M_{bc}(\mathcal{X})$ is the set of all compactly supported finite signed Borel measures on $\mathcal{X}$. Therefore, k being *cc*-universal is equivalent to the injectivity of (4.5)—in Remark 4.9, we present a more direct proof of this result. It is clear from the definitions of *c*- and *cc*-universality that these notions are equivalent when $\mathcal{X}$ is compact, which also follows from their characterizations in (4.2) and (4.5).

As special cases, Micchelli et al. [54, Propositions 14, 15 and Theorem 17] showed that a bounded continuous translation invariant kernel on $\mathbb{R}^d$, i.e., $k(x, y) = \psi(x - y)$ is *cc*-universal if $\mathrm{supp}(\Lambda)$ has a non-zero interior (a weaker condition of $\mathrm{supp}(\Lambda)$ being a uniqueness subset[10] of $\mathbb{C}^d$ is sufficient for a translation

[10] A subset S of $\mathbb{C}^d$ is a uniqueness set if an entire function (see footnote 4) on $\mathbb{C}^d$ vanishes on S then it is everywhere zero on $\mathbb{C}^d$.

invariant kernel on $\mathbb{R}^d$ to be *cc*-universal—see Proposition 14 of Micchelli et al., 2006), while a radial kernel on $\mathbb{R}^d$ is *cc*-universal if and only if $\text{supp}(\nu) \neq \{0\}$—see (2.2) and (2.5) for the definitions of Λ and ν. Using these characterizations, many popular kernels on $\mathbb{R}^d$ are shown to be *cc*-universal [54, Section 4]: Gaussian, Laplacian, B_{2l+1}-spline, sinc kernel, etc.

4.1.3 c_0- and L_p-universality

Although *cc*-universality solves the limitation of *c*-universality by handling non-compact $\mathcal{X}$, the topology of compact convergence considered in *cc*-universality is weaker than the topology of uniform convergence, i.e., a sequence of functions, $\{f_n\} \subset C(\mathcal{X})$ converging to $f \in C(\mathcal{X})$ in the topology of uniform convergence ensures that they converge in the topology of compact convergence but not vice-versa. So, the natural question to ask is whether we can characterize $\mathcal{H}$ that are rich enough to approximate any $f^\star$ on non-compact $\mathcal{X}$ in a stronger sense, i.e., uniformly, by some $g \in \mathcal{H}$. Carmeli et al. [13, Definition 2, Theorem 1] and Sriperumbudur et al. [77] answered this through the notion of c_0-universality, defined as follows.

Definition 4.3 (c_0-universality)**.** *A pd kernel, k is said to be a c_0-kernel if it is bounded with $k(\cdot, x) \in C_0(\mathcal{X}), \forall\, x \in \mathcal{X}$, where $\mathcal{X}$ is a locally compact Hausdorff (LCH) space. A c_0-kernel on an LCH space, $\mathcal{X}$ is said to be c_0-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $C_0(\mathcal{X})$ w.r.t. the supremum norm.*[11]

Note that a notion of universality that is stronger than c_0-universality can be defined by choosing $\mathcal{X}$ to be a Hausdorff space, $C_b(\mathcal{X})$ to be the target space and $\mathcal{H}$ being dense in $C_b(\mathcal{X})$ w.r.t. the supremum norm. However, this notion of universality does not enjoy a nice characterization as c_0-universality—see (4.6)

[11] Note that *cc*-universality (*resp.* *c*-universality) deals with $\mathcal{X}$ being a non-compact (*resp.* compact) Hausdorff space, whereas c_0-universality requires $\mathcal{X}$ to be an LCH space. While $\mathcal{X}$ being Hausdorff ensures that it has an abundance of compact subsets (as required in *cc*-universality), the stronger condition of $\mathcal{X}$ being an LCH space ensures that it has an abundance of continuous functions that vanish outside compact sets, which follows from Tietze extension theorem [26, Theorem 4.34]. In addition, this choice of $\mathcal{X}$ being an LCH space ensures the existence of topological dual of $C_0(\mathcal{X})$ through the Riesz representation theorem (Theorem C.3), which is required in the characterization of c_0-universality. See Proposition 4.8 for details.

and (4.7) for the characterization of c_0-universality—and therefore, we did not include it in our study of relationships between various notions of pd kernels. See Section 4.5 for details.

Before we present the characterization of c_0-universality, we need the definition of L_p-universality [81].

Definition 4.4 (L_p-universality)**.** *A measurable and bounded kernel, k defined on a Hausdorff space, $\mathcal{X}$ is said to be L_p-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $L^p(\mathcal{X},\mu)$ w.r.t. the L^p*-norm*, defined as $\|f\|_p := (\int_{\mathcal{X}} |f(x)|^p\,d\mu(x))^{1/p}$, for all Borel probability measures, μ, defined on $\mathcal{X}$ and some $p \in [1,\infty)$. Here $L^p(\mathcal{X},\mu)$ is the Banach space of p-integrable μ-measurable functions on $\mathcal{X}$.*

Carmeli et al. [13, Theorem 1] showed that a c_0-kernel k is c_0-universal if and only if it is L_p-universal, which by Theorem 2 and Proposition 3 of [13] is equivalent to the injectivity of the following embedding for all $\mu \in M_b(\mathcal{X})$ and some $p \in [1,\infty)$:

$$f \mapsto \int_{\mathcal{X}} k(\cdot,x)f(x)\,d\mu(x),\ f \in L^p(\mathcal{X},\mu). \tag{4.6}$$

We provide an alternate characterization for c_0-universality in Section 4.4 (see Proposition 4.8) that k is c_0-universal if and only if the following embedding is injective:

$$\mu \mapsto \int_{\mathcal{X}} k(\cdot,x)\,d\mu(x),\ \mu \in M_b(\mathcal{X}). \tag{4.7}$$

As a special case, [13, Proposition 16] showed that a bounded continuous translation invariant k on $\mathbb{R}^d$ is c_0-universal if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$. Examples of c_0-universal kernels on $\mathbb{R}^d$ include the Gaussian, Laplacian, B_{2l+1}-spline, inverse multiquadratics, Matérn class, etc.

4.1.4 Summary and Open Questions

The following statements summarize the relation between various notions of universality, which are depicted in Figure 4.1.

- c- and cc-universality are related to the injective RKHS embedding of finite signed Borel measures, as shown in (4.2) and (4.5).

- For an LCH space $\mathcal{X}$, c_0-universality implies cc-universality, which follows from (4.3) and (4.6). The converse is however not true as a translation invariant kernel on $\mathbb{R}^d$ is c_0-universal if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$ while $\text{int}(\text{supp}(\Lambda)) \neq \emptyset$ is sufficient for cc-universality.
- When $\mathcal{X}$ is compact, then c-, cc- and c_0-universality are equivalent.
- For an LCH space $\mathcal{X}$, a c_0-kernel is c_0-universal if and only if it is L_p-universal.
- If k is a radial kernel on $\mathbb{R}^d$, then k is cc-universal if and only if $\text{supp}(\nu) \neq \{0\}$.

The following relationships need to be clarified, which we do in Section 4.4.

(A) As mentioned in the summary, c- and cc-universality are related to the injective RKHS embedding of finite signed Borel measures. However, the relation between c_0-universality and the injective RKHS embedding of finite signed Borel measures as shown in (4.7) is not clear, which we clarify in Section 4.4.1.

(B) For an LCH space $\mathcal{X}$ (that is not compact), it is clear from the summary that c_0-universality implies cc-universality. Is there a case for which cc-universality implies c_0-universality? We address this in Section 4.4.3.

(C) While cc-universality is characterized for radial kernels on $\mathbb{R}^d$, the characterization of c_0-universality for radial kernels is not known. In Section 4.4.3, we provide a characterization of c_0-universality for radial kernels on $\mathbb{R}^d$ and then establish the relation between c_0-universality and cc-universality for such kernels.

4.2 Characteristic vs. Universal Kernels

In this section, we relate characteristic and universal kernels based on already existing characterizations for characteristic kernels and the results summarized in Section 4.1.4 for universal kernels.[12]

[12] While characteristic kernels ensure that $\mathbb{P} \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$ is injective where $\mathbb{P}$ is a Borel probability measure on a topological space $\mathcal{X}$, cc- and c_0-universal kernels ensure that (4.4) and (4.7) are injective where μ is a Radon measure. μ being a Radon measure follows from the Riesz representation theorem (Theorem C.3), which is used to obtain the measure embedding characterization for cc- and c_0-universality. Therefore, in order not to differentiate between Radon and Borel measures, we assumed $\mathcal{X}$ to be a Polish space (see Section C.1.1).

c-universal kernels vs. Characteristic kernels: [37] related universal and characteristic kernels by showing that if k is c-universal, then it is characteristic. The converse is not true: as an example, a bounded continuous translation invariant kernel, k on $\mathbb{T}^d \times \mathbb{T}^d$ is characteristic if and only if $A_\psi(0) \geq 0$, $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}^d_+$ while the following proposition shows that it is c-universal if and only if $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}^d$.

Proposition 4.5. *Suppose k satisfies Assumption 3.18. Then k is c-universal if and only $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}^d$, where A_ψ is defined in (2.4).*

Proof. $(\Leftarrow)$ Consider $\iint_{\mathbb{T}^d} k(x,y)\, d\mu(x)\, d\mu(y)$ for $\mu \in M_b(\mathbb{T}^d)\backslash\{0\}$. We have

$$\begin{aligned} B := \iint_{\mathbb{T}^d} k(x,y)\, d\mu(x)\, d\mu(y) &= \iint_{\mathbb{T}^d} \sum_{n\in\mathbb{Z}^d} A_\psi(n) e^{i\langle x-y,n\rangle}\, d\mu(x)\, d\mu(y) \\ &\overset{(a)}{=} \sum_{n\in\mathbb{Z}^d} A_\psi(n) \int_{\mathbb{T}^d} e^{i\langle x,n\rangle}\, d\mu(x) \int_{\mathbb{T}^d} e^{-i\langle y,n\rangle}\, d\mu(y) \\ &\overset{(b)}{=} (2\pi)^{2d} \sum_{n\in\mathbb{Z}^d} A_\psi(n) \overline{A_\mu(n)} A_\mu(n) \\ &= (2\pi)^{2d} \sum_{n\in\mathbb{Z}^d} A_\psi(n) |A_\mu(n)|^2, \end{aligned} \tag{4.8}$$

where Fubini's theorem (Theorem C.1) is invoked in (a) and

$$A_\mu(n) := (2\pi)^{-d} \int_{\mathbb{T}^d} e^{-i\langle n,x\rangle}\, d\mu(x),\ n \in \mathbb{Z}^d, \tag{4.9}$$

is used in (b). Note that A_μ is the Fourier transform of μ in $\mathbb{T}^d$. Since $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}^d$, we have $B > 0$, which means (4.2) is injective and therefore k is c-universal.

$(\Rightarrow)$ Proving necessity is equivalent to proving that if $A_\psi(n) = 0$ for some $n = n_0$, then there exists $\mu \in M_b(\mathbb{T}^d)\backslash\{0\}$ such that $\iint_{\mathbb{T}^d} k(x,y)\, d\mu(x)\, d\mu(y) = 0$.

Let $A_\psi(n) = 0$ for some $n = n_0$. Define $d\mu(x) = 2\alpha \cos(\langle x, n_0\rangle)\, dx$, $\alpha \in \mathbb{R}\backslash\{0\}$. By (4.9), we get $A_\mu(n) = \alpha\delta_{n_0,n}$, where δ represents the Kronecker delta. This means $\mu \neq 0$. Using A_ψ and A_μ in (4.8), it is easy to check that $\iint_{\mathbb{T}^d} k(x,y)\, d\mu(x)\, d\mu(y) = 0$, i.e., $\int_{\mathbb{T}^d} k(\cdot,x)\, d\mu = 0$, which means (4.2) is injective and therefore, k is not c-universal. □

cc-universal kernels vs. Characteristic kernels: cc-universal kernels on a non compact Hausdorff space need not be characteristic: for example, a translation invariant kernel on $\mathbb{R}^d$ is cc-universal if $\text{int}(\text{supp}(\Lambda)) \neq \emptyset$ (see the summary in

Section 4.1.4) while it is characteristic if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$ (see Theorem 3.13). Although, this example shows that a bounded continuous translation invariant kernel on $\mathbb{R}^d$ is *cc*-universal if it is characteristic, it is not clear whether such a relation holds on a general non-compact Hausdorff space (not necessarily $\mathbb{R}^d$). The following example shows that continuous kernels that are characteristic on non-compact Hausdorff space, $\mathcal{X}$ also need not be *cc*-universal.

Example 4.6. *Let $\mathcal{X} = \mathbb{N}$. Define $k(x,y) = \delta_{xy}$, $x, y \in \mathcal{X}\backslash\{1\}$, $k(x,1) = 0$ for any $x \in \mathcal{X}$, where δ represents the Kronecker delta. Suppose $\mu = \delta_1 \in M_{bc}(\mathcal{X})\backslash\{0\}$, where δ_j represents the Dirac measure at j. Then*

$$\left\|\int_{\mathcal{X}} k(\cdot,x)\,d\mu(x)\right\|_{\mathcal{H}}^2 = \|k(\cdot,1)\|_{\mathcal{H}}^2 = k(1,1) = 0,$$

which means there exists $\mu \in M_{bc}(\mathcal{X})\backslash\{0\}$ such that $\int_{\mathcal{X}} k(\cdot,x)\,d\mu(x) = 0$, i.e., (4.5) is not injective and therefore k is not cc-universal. However, k is characteristic as we show below.

Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $\mathcal{X}$ such that $\mathbb{P} = \sum_{j\in\mathbb{N}} p_j\delta_j$, $\mathbb{Q} = \sum_{j\in\mathbb{N}} q_j\delta_j$ with $p_j \geq 0, q_j \geq 0$ for all $j \in \mathbb{N}$ and $\sum_{j\in\mathbb{N}} p_j = \sum_{j\in\mathbb{N}} q_j = 1$. Consider

$$\begin{aligned}
B := \left\|\int_{\mathcal{X}} k(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(\mathcal{X})\right\|_{\mathcal{H}}^2 &= \left\|\sum_{j\in\mathbb{N}}(p_j - q_j)k(\cdot,j)\right\|_{\mathcal{H}}^2 \\
&= \sum_{l,j\in\mathbb{N}}(p_l - q_l)(p_j - q_j)k(l,j) \\
&= (p_1 - q_1)^2 k(1,1) + 2(p_1 - q_1)\sum_{j\in\mathbb{N}\backslash\{1\}}(p_j - q_j)k(j,1) \\
&\qquad + \sum_{l,j\in\mathbb{N}\backslash\{1\}}(p_j - q_j)(p_l - q_l)k(j,l) \\
&= \sum_{j\in\mathbb{N}\backslash\{1\}}(p_j - q_j)^2.
\end{aligned}$$

Suppose $B = 0$, which means $p_j = q_j$, $\forall\, j \in \mathbb{N}\backslash\{1\}$. Since $\sum_{j\in\mathbb{N}} p_j = \sum_{j\in\mathbb{N}} q_j = 1$, we have $p_1 = q_1$ and so $\mathbb{P} = \mathbb{Q}$, i.e., (3.1) is injective and therefore k is characteristic.

c_0-universal kernels vs. Characteristic kernels: [30, 31] have shown that a measurable and bounded kernel, k is characteristic if and only if the direct sum of $\mathcal{H}$

and $\mathbb{R}$ is dense in $L^p(\mathcal{X}, \mathbb{P})$ for all $\mathbb{P} \in M_+^1(\mathcal{X})$ and for some $p \in [1, \infty)$. Using this, it is easy to see that if $\mathcal{H}$ is dense in $L^p(\mathcal{X}, \mathbb{P})$ for all $\mathbb{P} \in M_+^1(\mathcal{X})$ and for some $p \in [1, \infty)$, then k is characteristic. Based on the results summarized in Section 4.1.4, it is clear that for an LCH space, $\mathcal{X}$, if k is c_0-universal, which means k is L_p-universal, then $\mathcal{H}$ is dense in $L^p(\mathcal{X}, \mathbb{P})$ for all $\mathbb{P} \in M_+^1(\mathcal{X})$ and for some $p \in [1, \infty)$ and therefore is characteristic. In Section 4.4, we provide an alternate proof for this relation between c_0-universal and characteristic kernels by answering (A). Clearly, the converse is not true, i.e., a c_0-kernel that is characteristic need not be c_0-universal (see Proposition 4.10 and footnote 14). However, for bounded continuous translation invariant kernels on $\mathbb{R}^d$, the converse is true, i.e., a c_0-kernel that is characteristic is also c_0-universal. This is because of the fact that a translation invariant kernel on $\mathbb{R}^d$ is characteristic if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$ (see Theorem 3.13), which is also the same characterization summarized in Section 4.1.4 for c_0-universal kernels.

Summary: The following statements summarize the relation between universal and characteristic kernels, which are depicted in Figure 4.1.

- For c_0-kernels defined on an LCH space, $\mathcal{X}$, L_p-universal $\Leftrightarrow$ c_0-universal $\Rightarrow$ characteristic. But in general, c_0-kernels that are characteristic need not be c_0-universal. However, for bounded continuous translation invariant kernels on $\mathbb{R}^d$, c_0-universal $\Leftrightarrow$ characteristic.
- When $\mathcal{X}$ is compact, c-universal $\Rightarrow$ characteristic but not vice-versa.
- For bounded continuous translation invariant kernels on $\mathbb{R}^d$, characteristic $\Rightarrow$ cc-universal but not vice-versa. However, on general non-compact Hausdorff spaces, continuous kernels that are characteristic need not be cc-universal.

Open questions: The following relationship need to be clarified, which we do in Section 4.4.

(D) While the relation between universal and characteristic kernels that are translation invariant on $\mathbb{R}^d$ is clear (see the summary above), the characterization of c_0-universal kernels that are radial on $\mathbb{R}^d$ is not known and therefore the relation between characteristic and universal kernels that are radial on $\mathbb{R}^d$ is not clear. We address this in Section 4.4.3.

4.3 Universal & Characteristic Kernels vs. Others

In this section, we relate characteristic kernels and various notions of universal kernels to strictly pd, integrally strictly pd and conditionally strictly pd kernels. Before that, we summarize the relation between strictly pd, integrally strictly pd and conditionally strictly pd kernels. While integrally strictly pd kernels are strictly pd (see Proposition 2.6), the converse is not true, which follows from [81, Proposition 4.60, Theorem 4.62]. However, if $\mathcal{X}$ is a finite set, then k being strictly pd also implies it is integrally strictly pd. From the definitions of strictly pd and conditionally strictly pd kernels, it is clear that a strictly pd kernel is conditionally strictly pd but not vice-versa.

Universal kernels vs. Strictly pd kernels: [13, Corollary 5] showed that cc-universal kernels are strictly pd, which means c_0-universal kernels are also strictly pd (as c_0-universal $\Rightarrow$ cc-universal from Section 4.1.4). This means, when $\mathcal{X}$ is compact Hausdorff, c-universal kernels are strictly pd, which matches with the result in [81, Definition 4.53, Proposition 4.54, Example 4.11].

Conversely, a strictly pd c_0-kernel on an LCH space need not be c_0-universal. This follows from Theorem 4.62 in [81] which shows that there exists a bounded strictly pd kernel, k on $\mathcal{X} := \mathbb{N} \cup \{0\}$ with $k(\cdot, x) \in C_0(\mathcal{X})$, $\forall\, x \in \mathcal{X}$ such that k is not L_p-universal (which from Section 4.1.4 means k is not c_0-universal). Similarly, when $\mathcal{X}$ is compact, the converse is not true, i.e., continuous strictly pd kernels need not be c-universal which follows from the results due to [17] and [58] for Taylor kernels [81, Lemma 4.8, Corollary 4.57]—refer to [81, Section 4.7, p. 161] for more details.[13] While it is not evident whether a continuous strictly pd kernel is in gen-

[13] Another example of continuous strictly pd kernels that are not c-universal is as follows. By Proposition 4.5, a bounded continuous translation invariant kernel on $\mathbb{T} \times \mathbb{T}$ is c-universal if and only if $A_\psi(n) > 0$, $\forall\, n \in \mathbb{Z}$. Therefore, by Theorem C.6, a strictly pd kernel on $\mathbb{T}$ need not be c-universal.

eral *cc*-universal or not, it is indeed the case for translation invariant kernels that are continuous, bounded and integrable on $\mathbb{R}^d$, i.e., $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$, where $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$. This follows from Theorem 6.11 and Corollary 6.12 of [91] that if $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ is strictly pd, then $\text{int}(\text{supp}(\Lambda)) \neq \emptyset$, which from Section 4.1.4 means k is *cc*-universal. Similarly, when the kernel is radial on $\mathbb{R}^d$, then strictly pd kernels are *cc*-universal. This follows from Theorem 7.14 of [91], which shows that a radial kernel on $\mathbb{R}^d$ is strictly pd if and only if $\text{supp}(\nu) \neq \{0\}$, and therefore *cc*-universal (from Section 4.1.4). On the other hand, when $\mathcal{X}$ is finite, all these notions of universal and strictly pd kernels are equivalent, which follows from the result due to Carmeli et al. [13, Corollary 5] that *cc*-universal and strictly pd kernels are the same when $\mathcal{X}$ is finite.

Characteristic kernels vs. Strictly pd kernels: Since bounded continuous translation invariant characteristic kernels on $\mathbb{R}^d$ are equivalent to c_0-universal kernels (see Section 4.2), it is clear that they are strictly pd. However, the converse is not true: for example, the sinc-squared kernel, which has $\text{supp}(\Lambda) = [-\sigma, \sigma]^d \subsetneq \mathbb{R}^d$ is strictly pd [91, Theorem 6.11], while it is not characteristic. However, the converse is true for radial kernels on $\mathbb{R}^d$, which follows from Corollary 3.16 and [91, Theorem 7.14]. Based on Example 4.6, it can be shown that in general, characteristic kernels on a non-compact space (not necessarily $\mathbb{R}^d$) need not be strictly pd: in Example 4.6, k is characteristic but is not strictly pd because for $(a_1, \ldots, a_n) = (1, 0, \ldots, 0)$ and $(x_1, \ldots, x_n) = (1, \ldots, n)$, we have $\sum_{l,j=1}^n a_l a_j k(x_l, x_j) = a_1^2 k(1,1) + 2a_1 \sum_{j=2}^n a_j k(j,1) + \sum_{j=2}^n a_j^2 = 0$. Note that Example 4.6 holds even if $\mathcal{X}$ is a compact subset of $\mathbb{N}$. Therefore, when $\mathcal{X}$ is compact Hausdorff, a characteristic kernel need not be strictly pd. However, for bounded continuous translation invariant kernels on $\mathbb{T}$, a characteristic kernel is also strictly pd, while the converse is not true: Theorem 3.19 (also see [31, Theorem 8]) shows that k on $\mathbb{T} \times \mathbb{T}$ is characteristic if and only if $A_\psi(0) \geq 0$, $A_\psi(n) > 0$, $\forall\, n \in \mathbb{Z}\backslash\{0\}$, which by Theorem C.6 is strictly pd, while the converse is clearly not true.

Characteristic kernels vs. Integrally strictly pd kernels: Theorem 3.10 shows that integrally strictly pd kernels are characteristic, while the converse in general is not true.[14] When k is translation invariant on $\mathbb{R}^d$, however the converse holds, which is due to the fact that if k is characteristic, then $\text{supp}(\Lambda) = \mathbb{R}^d$ (see Theorem 3.13), which ensures that k is integrally strictly pd.

Summary: The following statements summarize the relation of universal and characteristic kernels to strictly pd kernels, integrally strictly pd and conditionally strictly pd, which are depicted in Figure 4.1.

- c-, cc- and c_0-universal kernels are strictly pd and are therefore conditionally strictly pd, while the converse is not true in general. When $\mathcal{X}$ is finite, then c-, cc- and c_0-universal kernels are equivalent to strictly pd kernels.
- Bounded, continuous, integrable, strictly pd translation invariant kernels on $\mathbb{R}^d$ are cc-universal. Radial kernels on $\mathbb{R}^d$ are strictly pd if and only if they are cc-universal.
- For a general non-compact Hausdorff space, characteristic kernels need not be strictly pd and vice-versa. However, bounded continuous translation invariant kernels on $\mathbb{R}^d$ or $\mathbb{T}$ that are characteristic are strictly pd but the converse is not true. The converse also holds if k is radial on $\mathbb{R}^d$.
- Integrally strictly pd kernels are characteristic. Though the converse is not true in general, it holds if the kernel is bounded, continuous and translation invariant on $\mathbb{R}^d$.

Open questions: The following questions need to be clarified, which is done in Section 4.4.

(E) While the relation of universal kernels to strictly pd and conditionally strictly pd kernels is clear from the above summary, the relation between universal and integrally strictly pd kernels is not known, which we establish in Section 4.4.2.

(F) When $\mathcal{X}$ is a finite set, it is easy to see that characteristic and conditionally strictly pd kernels are equivalent (see Section 4.4.4). However, their

[14] By Example 4.6, it is clear that for $\mu = \delta_1 \in M_b(\mathcal{X})\backslash\{0\}$, $\iint_{\mathcal{X}} k(x,y)\, d\mu(x)\, d\mu(y) = k(1,1) = 0$, where δ_1 represents the Dirac measure at 1. Therefore k is not integrally strictly pd but is characteristic.

relationship is not clear for a general measurable space, which we clarify in Section 4.4.4.

(G) As summarized above, radial kernels on $\mathbb{R}^d$ are strictly pd if and only if they are *cc*-universal (which are also characteristic). However, the relation between all the other notions of pd kernels—c_0-universal, strictly pd and integrally strictly pd—is not known, which is addressed in Section 4.4.3.

4.4 New Results

In this section, we address the open questions, (A)–(G) mentioned in Sections 4.1–4.3 to understand the complete relationship between various notions of pd kernels.

4.4.1 c_0-universality and RKHS Embedding of Measures

As mentioned in Section 4.1, Micchelli et al. [54] have established the relation of *c*-universality and *cc*-universality to injective RKHS embedding of finite signed Borel measures (shown in (4.2) and (4.5)) through a simple application of the Hahn-Banach theorem (see Theorem 4.7). The following result in Proposition 4.8 provides a measure embedding characterization (shown in (4.7)) for c_0-universality, which is also obtained as a simple application of the Hahn-Banach theorem, and therefore addresses the open question in (A). Before we state Proposition 4.8, we present the Hahn-Banach theorem, which we quote from [65, Theorem 3.5 and the remark following Theorem 3.5].

Theorem 4.7 (Hahn-Banach)**.** *Suppose A is a subspace of a locally convex topological vector space Y. Then A is dense in Y if and only if $A^\perp = \{0\}$, where*

$$A^\perp := \{T \in Y' : \forall x \in A,\ T(x) = 0\}.$$

Here Y' denotes the topological dual of Y.

The following result, which presents a necessary and sufficient condition for k to be c_0-universal hinges on the above theorem, where we choose A to be the

RKHS, $\mathcal{H}$ and Y to be $C_0(\mathcal{X})$ for which Y' is known through the Riesz representation theorem.

Proposition 4.8 (c_0-universality and RKHS embedding of measures)**.** *Suppose $\mathcal{X}$ is an LCH space with the kernel, k being bounded and $k(\cdot,x)\in C_0(\mathcal{X})$, $\forall\, x\in\mathcal{X}$. Then k is c_0-universal if and only if the embedding,*

$$\mu\mapsto\int_{\mathcal{X}}k(\cdot,x)\,d\mu(x),\;\;\mu\in M_b(\mathcal{X}),\tag{4.10}$$

is injective.

Proof. By definition, k is c_0-universal if $\mathcal{H}$ is dense in $C_0(\mathcal{X})$. We now invoke Theorem 4.7 to characterize the denseness of $\mathcal{H}$ in $C_0(\mathcal{X})$, which means we need to consider the dual $C'_0(\mathcal{X}) := (C_0(\mathcal{X}))'$ of $C_0(\mathcal{X})$. By the Riesz representation theorem [26, Theorem 7.17], $C'_0(\mathcal{X}) = M_b(\mathcal{X})$ in the sense that there is a bijective linear isometry $\mu\mapsto T_\mu$ from $M_b(\mathcal{X})$ onto $C'_0(\mathcal{X})$, given by the natural mapping, $T_\mu(f)=\int_{\mathcal{X}}f\,d\mu,\;f\in C_0(\mathcal{X})$. Therefore, by Theorem 4.7, $\mathcal{H}$ is dense in $C_0(\mathcal{X})$ if and only if $\mathcal{H}^\perp := \{\mu\in M_b(\mathcal{X}) : \forall\, f\in\mathcal{H},\;\int_{\mathcal{X}}f\,d\mu=0\}=\{0\}$. From Lemma 3.1, we have $\mathcal{H}^\perp=\{\mu\in M_b(\mathcal{X}) : \int_{\mathcal{X}}k(\cdot,x)\,d\mu(x)=0\}$ and therefore the result follows from Theorem 4.7. □

Remark 4.9. (*a*) *When $\mathcal{X}$ is compact, $C_0(\mathcal{X})$ coincides with $C(\mathcal{X})$, and therefore the result in (4.10) matches with the one in (4.2), derived by [54].*

(*b*) *The characterization of cc-universality, shown in (4.5) can also be directly obtained as a simple application of Theorem 4.7, wherein the proof is similar to that of Proposition 4.8 except that we need to consider the dual of $C(\mathcal{X})$ endowed with the topology of compact convergence (a locally convex topological vector space) to characterize the denseness of $\mathcal{H}$ in $C(\mathcal{X})$. It is known [42] that $C'(\mathcal{X}) = M_{bc}(\mathcal{X})$ in the sense that there is a bijective linear isometry $\mu\mapsto T_\mu$ from $M_{bc}(\mathcal{X})$ onto $C'(\mathcal{X})$, given by the natural mapping, $T_\mu(f)=\int_{\mathcal{X}}f\,d\mu,\;f\in C(\mathcal{X})$. The rest of the proof is verbatim with $M_b(\mathcal{X})$ replaced by $M_{bc}(\mathcal{X})$.*

(*c*) *Comparing (4.10) and (3.1), it is clear that c_0-universal kernels are characteristic while the converse is not true, which matches with the result in Section 4.2.*

4.4.2 Universal vs. Integrally Strictly Positive Definite Kernels

In this section, we address the open question (E) through the following result which shows that k is c_0-universal if and only if it is integrally strictly pd.

Proposition 4.10 (c_0-universal and integrally strictly pd kernels)**.** *Suppose the assumptions in Proposition 4.8 hold. Then, k is c_0-universal if and only if it is integrally strictly pd.*

Proof. ($\Leftarrow$) Suppose k is not c_0-universal. By Proposition 4.8, there exists $\mu \in M_b(\mathcal{X})\backslash\{0\}$ such that $\int_{\mathcal{X}} k(\cdot, x)\, d\mu(x) = 0$, which implies $\| \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x)\|_{\mathcal{H}} = 0$. This means

$$0 = \left\langle \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x), \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x) \right\rangle_{\mathcal{H}} \overset{(e)}{=} \iint_{\mathcal{X}} k(x, y)\, d\mu(x)\, d\mu(y),$$

i.e., k is not integrally strictly pd, where (e) follows from Lemma 3.1.

($\Rightarrow$) Suppose k is not integrally strictly pd, i.e., there exists $\mu \in M_b(\mathcal{X})\backslash\{0\}$ such that $\iint_{\mathcal{X}} k(x, y)\, d\mu(x)\, d\mu(y) = 0$, i.e., $\| \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x)\|_{\mathcal{H}} = 0$, which implies $\int_{\mathcal{X}} k(\cdot, x)\, d\mu(x) = 0$. Therefore, the embedding in (4.10) is not injective, which by Proposition 4.8 implies that k is not c_0-universal. □

4.4.3 Radial kernels on $\mathbb{R}^d$

In this section, we address the open questions (B), (C), (D) and (G) by showing that all the notions of universality and characteristic kernels are equivalent to strictly pd kernels.

Proposition 4.11 (All notions are equivalent for radial kernels on $\mathbb{R}^d$)**.** *Suppose k is radial on $\mathbb{R}^d$. Then the following conditions are equivalent.*

(a) $\text{supp}(\nu) \neq \{0\}$.

(b) k is integrally strictly pd.

(c) k is c_0-universal.

(d) k is cc-universal.

(e) k is strictly pd.

(f) k is characteristic.

Proof. Note that $(b) \Leftrightarrow (c) \Rightarrow (d) \Leftrightarrow (e)$ which follows from Proposition 4.10 and results summarized in Sections 4.1.4 and 4.3. Theorem 7.14 in [91] ensures that $(e) \Rightarrow (a)$. $(a) \Rightarrow (f)$ follows from Corollary 3.16. Now, we show $(a) \Rightarrow (b)$.

Consider $\iint_{\mathbb{R}^d} k(x,y)\,d\mu(x)\,d\mu(y)$ with k as in (2.5), given by

$$\begin{aligned}
B &:= \iint_{\mathbb{R}^d} k(x,y)\,d\mu(x)\,d\mu(y)\\
&= \iint_{\mathbb{R}^d} \int_0^\infty e^{-t\|x-y\|_2^2}\,d\nu(t)\,d\mu(x)\,d\mu(y)\\
&\overset{(\star)}{=} \int_0^\infty \left(\iint_{\mathbb{R}^d} e^{-t\|x-y\|_2^2}\,d\mu(x)\,d\mu(y)\right)d\nu(t)\\
&\overset{(\clubsuit)}{=} \int_0^\infty \frac{1}{(4\pi t)^{d/2}}\left(\int_{\mathbb{R}^d} |\widehat{\mu}(\omega)|^2 e^{-\frac{\|\omega\|_2^2}{4t}}\,d\omega\right)d\nu(t)\\
&\overset{(\spadesuit)}{=} \int_{\mathbb{R}^d} |\widehat{\mu}(\omega)|^2 \left(\int_0^\infty \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|\omega\|_2^2}{4t}}\,d\nu(t)\right)d\omega, \qquad (4.11)
\end{aligned}$$

where Fubini's theorem (Theorem C.1) is invoked in $(\star)$ and $(\spadesuit)$, while we used (2.7) in $(\clubsuit)$, where we set $\psi(x) = e^{-t\|x\|_2^2}$ with $d\Lambda(\omega) = (4\pi t)^{-d/2}e^{-\|\omega\|_2^2/4t}\,d\omega$. Since $\text{supp}(\nu) \neq \{0\}$, the inner integral in (4.11) is positive for every $\omega \in \mathbb{R}^d$ and so $B > 0$, which means k is integrally strictly pd. □

4.4.4 Characteristic vs. Conditionally Strictly pd Kernels

In this section we address the open question (F) which is about the relation of characteristic kernels to conditionally strictly pd kernels. As shown in Section 4.3, although the relation between universal and conditionally strictly pd kernels straightforwardly follows from universal kernels being strictly pd, which in turn are conditionally strictly pd, such an implication is not possible in the case of characteristic kernels as they are not in general strictly pd (see Example 4.6). However, the following result establishes the relation between characteristic and conditionally strictly pd kernels.

Proposition 4.12. *If k is characteristic, then it is conditionally strictly pd.*

Proof. Suppose k is not conditionally strictly pd. This means for some $n \geq 2$ and for mutually distinct $x_1, \ldots, x_n \in \mathcal{X}$, there exists $\{\alpha_j\}_{j=1}^n \neq 0$ with $\sum_{j=1}^n \alpha_j = 0$ such that $\sum_{l,j=1}^n \alpha_l \alpha_j k(x_l, x_j) = 0$. Define $\mathcal{I} := \{j : \alpha_j > 0\}$, $\mathbb{P} := \beta^{-1} \sum_{j \in \mathcal{I}} \alpha_j \delta_{x_j}$ and $\mathbb{Q} := -\beta^{-1} \sum_{j \notin \mathcal{I}} \alpha_j \delta_{x_j}$, where $\beta := \sum_{j \in \mathcal{I}} \alpha_j$. It is easy to see that $\mathbb{P}$ and $\mathbb{Q}$ are distinct Borel probability measures on $\mathcal{X}$. Then, we have

$$\left\| \int_{\mathcal{X}} k(\cdot, x)\, d(\mathbb{P} - \mathbb{Q})(x) \right\|_{\mathcal{H}}^2 = \beta^{-2} \left\| \sum_{j=1}^n \alpha_j k(\cdot, x_j) \right\|_{\mathcal{H}} = \beta^{-2} \sum_{l,j=1}^n \alpha_l \alpha_j k(x_l, x_j) = 0.$$

So, there exist $\mathbb{P} \neq \mathbb{Q}$ such that $\int_{\mathcal{X}} k(\cdot, x)\, d(\mathbb{P} - \mathbb{Q})(x) = 0$, i.e., k is not characteristic. □

The converse to Proposition 4.12 in general is however not true: we showed in Section 4.3 that strictly pd kernels are conditionally strictly pd but need not be characteristic and so conditionally strictly pd kernels need not have to be characteristic. In the following, we present a concrete example to show the same—a similar example is used to prove Theorem 4.62 in [81], which shows that c_0-kernels that are strictly pd need not be c_0-universal.

Example 4.13. *Let $\mathcal{X} = \mathbb{N} \cup \{0\}$. Define $k(0,0) = \sum_{n \in \mathbb{N}} b_n^2$, $k(m,n) = \delta_{mn}$ and $k(n,0) = b_n$ for $m, n \geq 1$, where $\{b_n\}_{n \geq 1} \subset (0,1)$ and $\sum_{n \in \mathbb{N}} b_n = 1$. Let $n \geq 2$ and $\alpha := (\alpha_0, \ldots, \alpha_n) \in \mathbb{R}^{n+1}$ be a vector with $\alpha \neq 0$ such that $\sum_{j=0}^n \alpha_j = 0$. Consider*

$$\begin{aligned} B := \sum_{l,j=0}^n \alpha_l \alpha_j k(l,j) &= \alpha_0^2 k(0,0) + 2\sum_{j=1}^n \alpha_j \alpha_0 k(j,0) + \sum_{l,j=1}^n \alpha_l \alpha_j k(l,j) \\ &= \alpha_0^2 \sum_{j \in \mathbb{N}} b_j^2 + 2\alpha_0 \sum_{j=1}^n \alpha_j b_j + \sum_{j=1}^n \alpha_j^2 \\ &= = \alpha_0^2 \sum_{j \in \mathbb{N}} b_j^2 + \sum_{j=1}^n \alpha_j (2\alpha_0 b_j + \alpha_j). \end{aligned}$$

If $\alpha_0 = 0$, then $B = \sum_{j=1}^n \alpha_j^2 > 0$ since we assumed $\alpha \neq 0$. Suppose $\alpha_0 \neq 0$. Then

$$B \geq \alpha_0^2 \sum_{j \in \mathbb{N}} b_j^2 + \sum_{j=1}^n \alpha_j^* (2\alpha_0 b_j + \alpha_j^*), \tag{4.12}$$

where

$$(\alpha_1^*,\ldots,\alpha_n^*)=\arg\min\left\{\sum_{j=1}^n\alpha_j(2\alpha_0 b_j+\alpha_j)\,:\,\sum_{j=1}^n\alpha_j=-\alpha_0\right\}. \tag{4.13}$$

Note that $(\alpha_1^*,\ldots,\alpha_n^*)$ *is unique as the objective in (4.13) is strictly convex, which is minimized over a convex set. To solve (4.13), let us consider the Lagrangian, given as*

$$L(\alpha_1,\ldots,\alpha_n,\lambda)=\sum_{j=1}^n\alpha_j(2\alpha_0 b_j+\alpha_j)-\lambda\Big(\sum_{j=1}^n\alpha_j+\alpha_0\Big),$$

where $\lambda\ge 0$. *Differentiating* L *w.r.t.* α_j *and setting it to zero yields* $\alpha_j^*=(\lambda-2\alpha_0 b_j)/2$. *Since* $\sum_{j=1}^n\alpha_j^*=-\alpha_0$, *we have* $\lambda=\frac{2\alpha_0(a-1)}{n}$, *where* $a:=\sum_{j=1}^n b_j$. *Substituting for* λ *in* α_j^*, *we have*

$$\alpha_j^*=\frac{\alpha_0(a-1-nb_j)}{n},\ j\in\mathbb{N}_n.$$

Substituting for α_j^* *in (4.12) gives*

$$B\ge\alpha_0^2\sum_{j\in\mathbb{N}}b_j^2+\frac{\alpha_0^2(a-1)^2}{n}-\alpha_0^2\sum_{j=1}^n b_j^2=\alpha_0^2\sum_{j=n+1}^\infty b_j^2+\frac{\alpha_0^2(\sum_{j=1}^n b_j-1)^2}{n}>0.$$

Consequently, we have $B>0$ *in any case, and therefore* k *is conditionally strictly pd. In the following, we however show that* k *is not characteristic.*

Let $\mathbb{P}=\delta_0$ *and* $\mathbb{Q}=\sum_{j=1}^n b_j\delta_j$. *Clearly* $\mathbb{P}\ne\mathbb{Q}$. *Consider*

$$\begin{aligned}\Big\|\int_{\mathcal{X}}k(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\Big\|_{\mathcal{H}}^2&=\Big\|k(\cdot,0)-\sum_{j\in\mathbb{N}}k(\cdot,j)b_j\Big\|_{\mathcal{H}}^2\\&=k(0,0)-2\sum_{j\in\mathbb{N}}k(j,0)b_j+\sum_{l,j\in\mathbb{N}}k(l,j)b_lb_j\\&=\sum_{j\in\mathbb{N}}b_j^2-2\sum_{j\in\mathbb{N}}b_j^2+\sum_{j\in\mathbb{N}}b_j^2=0.\end{aligned}$$

This implies (3.1) is not injective and therefore k *is not characteristic.*

When $\mathcal{X}$ is finite, then the converse to Proposition 4.12 holds, i.e., conditionally strictly pd kernels are characteristic, which is shown as follows. Let $\mathcal{X}=\mathbb{N}_n$. Suppose k is conditionally strictly pd, i.e., for any $n\ge 2$, $(\alpha_1,\ldots,\alpha_n)\ne(0,\ldots,0)$

with $\sum_{j=1}^{n} \alpha_j = 0$, and all distinct $x_1, \ldots, x_n \in \mathcal{X}$, we have $\sum_{l,j=1}^{n} \alpha_l \alpha_j k(x_l, x_j) > 0$. Let $\mathcal{I} := \{j : \alpha_j > 0\}$. Define $\mathbb{P} := \beta^{-1} \sum_{j \in \mathcal{I}} \alpha_j \delta_j$ and $\mathbb{Q} := -\beta^{-1} \sum_{j \notin \mathcal{I}} \alpha_j \delta_j$, where $\beta := \sum_{j \in \mathcal{I}} \alpha_j$ and $\mathbb{P} \neq \mathbb{Q}$. Then

$$\left\| \int_{\mathcal{X}} k(\cdot, x)\, d(\mathbb{P} - \mathbb{Q})(x) \right\|_{\mathcal{H}}^{2} = \beta^{-2} \sum_{l,j=1}^{n} \alpha_l \alpha_j k(l, j) > 0$$

and therefore k is characteristic.

4.5 c_b-universality

As mentioned in Section 4.1.3, the definition of c_0-universality deals with $\mathcal{H}$ being dense in $C_0(\mathcal{X})$ w.r.t. the supremum norm, where $\mathcal{X}$ is an LCH space. Although the notion of c_0-universality addresses limitations associated with both c- and cc-universality, it only approximates a subset of $C(\mathcal{X})$, i.e., it cannot deal with functions in $C(\mathcal{X}) \backslash C_0(\mathcal{X})$. This limitation can be addressed by considering a larger class of functions to be approximated.

To this end, one can consider a notion of universality that is stronger than c_0-universality: k is said to be c_b-universal if its corresponding RKHS, $\mathcal{H}$ is dense in $C_b(\mathcal{X})$, the space of bounded continuous functions on a topological space, $\mathcal{X}$ (note that $C_0(\mathcal{X}) \subset C_b(\mathcal{X})$). This notion of c_b-universality may be more applicable in learning theory than c_0-universality as the target function, $f^\star$ can belong to $C_b(\mathcal{X})$ (which is a more natural assumption) instead of it being restrained to $C_0(\mathcal{X})$ (note that $C_0(\mathcal{X})$ only contains functions that vanish at infinity). Similar to Proposition 4.8, the following theorem provides a necessary and sufficient condition for k to be c_b-universal. Before we state the result, we need some definitions.

A *set function* is a function defined on a family of sets, and has values in $[-\infty, +\infty]$. A set function μ defined on a family τ of sets is said to be *finitely additive* if $\emptyset \in \tau$, $\mu(\emptyset) = 0$ and $\mu(\cup_{l=1}^{n} A_l) = \sum_{l=1}^{n} \mu(A_l)$, for every finite family $\{A_1, \ldots, A_n\}$ of disjoint subsets of τ such that $\cup_{l=1}^{n} A_l \in \tau$. A *field of subsets* of a set $\mathcal{X}$ is a non-empty family, Σ, of subsets of $\mathcal{X}$ such that $\emptyset \in \Sigma$, $\mathcal{X} \in \Sigma$, and for all $A, B \in \Sigma$, we have $A \cup B \in \Sigma$ and $B \backslash A \in \Sigma$. An additive set function μ defined on a field Σ of subsets of a topological space $\mathcal{X}$ is said to be *regular* if for each $A \in \Sigma$ and $\epsilon > 0$, there exists $B \in \Sigma$ whose closure is contained in A and there exists $C \in \Sigma$ whose interior contains A such that $|\mu(D)| < \epsilon$ for every $D \in \Sigma$ with $D := C\backslash B$.

Proposition 4.14 (c_b-universality and RKHS embedding of set functions)**.** *Suppose $\mathcal{X}$ is a normal topological space and $M_{rba}(\mathcal{X})$ is the space of all finitely additive, regular, bounded set functions defined on the field generated by the closed sets of $\mathcal{X}$. Then, a bounded continuous kernel, k is c_b-universal if and only if the embedding,*

$$\mu \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mu,\ \mu \in M_{rba}(\mathcal{X}), \tag{4.14}$$

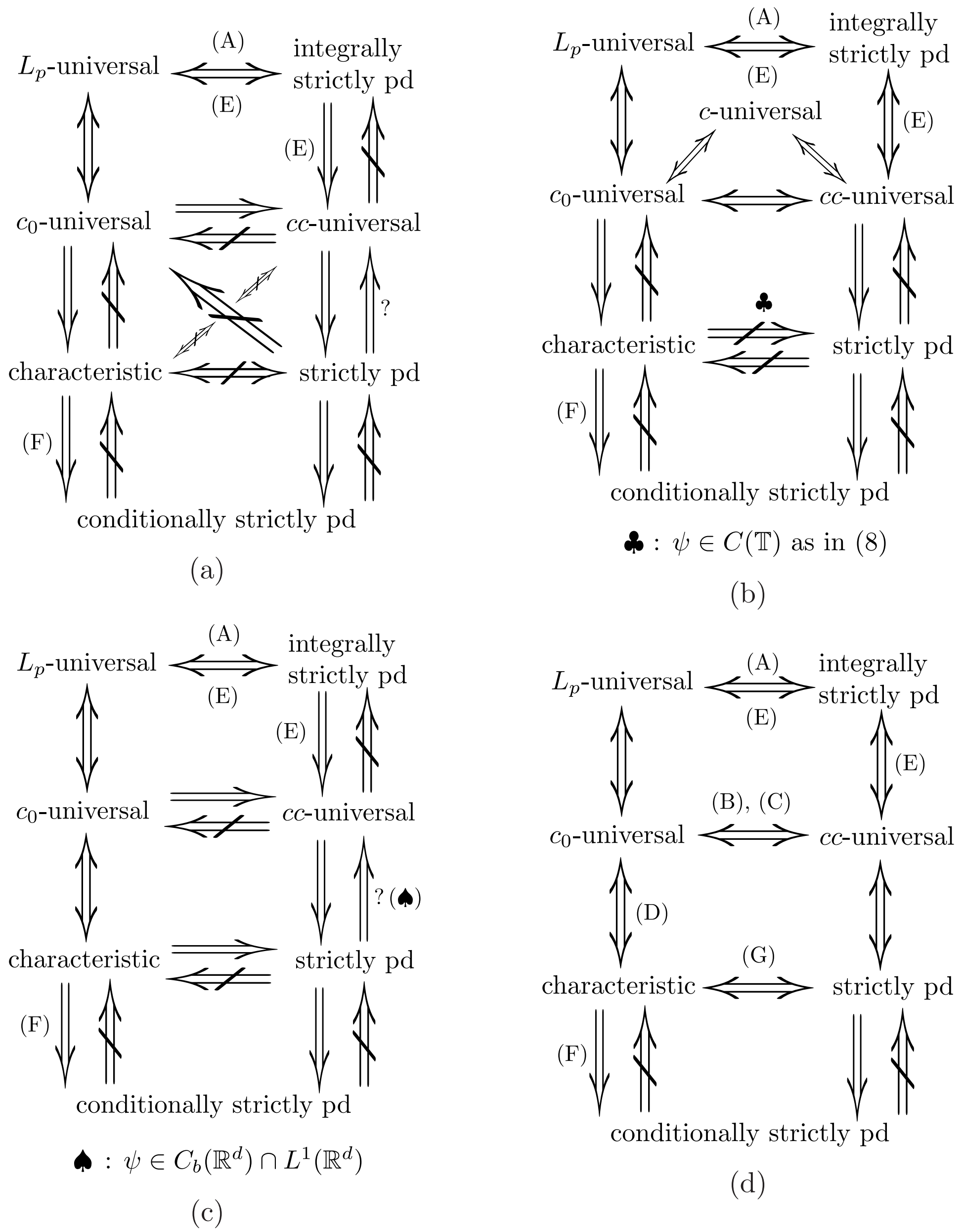
is injective.

Proof. The proof is very similar to that of Proposition 4.8, wherein we identify $(C_b(\mathcal{X}))' \cong M_{rba}(\mathcal{X})$ such that $T \in (C_b(\mathcal{X}))'$ and $\mu \in M_{rba}(\mathcal{X})$ satisfy $T(f) = \int_{\mathcal{X}} f\, d\mu$, $f \in C_b(\mathcal{X})$ [24, p. 262]. Here, $\cong$ represents the isometric isomorphism. The rest of the proof is verbatim with $M_bf(\mathcal{X})$ replaced by $M_{rba}(\mathcal{X})$. □

Note that $M_{rba}(\mathcal{X})$ does not contain any measure—though a set function in $M_{rba}(\mathcal{X})$ can be extended to a measure—as measures are countably additive and defined on a σ-field. Since μ in Proposition 4.14 is not a measure but a finitely additive set function defined on a field, it is not clear how to deal with the integral in (4.14). Because of the technicalities involved in dealing with set functions, the analysis of c_b-universality and its relation to other notions considered in Sections 4.1–4.3 is not clear, although it is an interesting problem to be resolved because of its applicability in learning theory.

Bibliographic Notes

This chapter is based on joint work with Kenji Fukumizu and Gert Lanckriet, which appeared in [77]. The longer version of [77] is currently under submission to the Journal of Machine Learning Research. The dissertation author was the primary investigator and author of these papers.



♣ : $\psi \in C(\mathbb{T})$ as in (8)

♠ : $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$

Figure 4.1: Summary of the relations between various families of kernels: The implications shown without any reference are based on the review of existing results (see Sections 4.1–4.3) while the ones with a reference are based on new results derived in Section 4.4 that addresses the open questions (A)–(G). The implications which are still open are shown with "?". (a) $\mathcal{X}$ is an LCH space. (b) The implications shown hold for any compact Hausdorff space, $\mathcal{X}$. When $\mathcal{X} = \mathbb{T}$ and k is translation invariant on $\mathbb{T}$ (see (2.4)), then k being characteristic implies it is strictly pd (spd), which is shown as ♣. (c) The implications shown hold for translation invariant kernels on $\mathbb{R}^d$ (see (2.2)). If $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$, then the implication shown as (♠) holds, i.e., spd kernels are cc-universal. Otherwise, it is not clear whether the implication holds. (d) Radial kernels on $\mathbb{R}^d$ (see (2.5)).

5 Integral Probability Metrics, ϕ-Divergences and MMD

In Chapters 3 and 4, we discussed the question of when is k characteristic so that γ_k (i.e., MMD) is a metric on the space of probability measures. Many distance measures on probabilities have been studied in literature, of which two popular families are: (i) Integral probability metrics and (ii) ϕ-divergences. The goal of this chapter is to study the relation of MMD to these families, in particular the advantages and disadvantages of MMD over these families. We briefly discussed the relation between MMD and IPMs in Chapter 3 (see Proposition 3.2), wherein we showed that MMD is obtained by choosing $\mathcal{F} = \mathcal{F}_k := \{f : \|f\|_{\mathcal{H}} \leq 1\}$ in (3.2). In this chapter, we elaborate on this result by considering the question of "What are the advantages of choosing $\mathcal{F} = \mathcal{F}_k$ in (3.2) compared to the other choices of $\mathcal{F}$?"

5.1 Introduction

The notion of distance between probability measures has found many applications in probability theory and mathematical statistics. One of the widely studied and well understood family of distance measures (in fact divergences) between probabilities is the *Ali-Silvey distance* [1], also called the *Csiszár's ϕ-divergence* [15], which is defined as

$$D_\phi(\mathbb{P}, \mathbb{Q}) := \int_{\mathcal{X}} \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \text{ if } \mathbb{P} \ll \mathbb{Q},$$

where $\mathcal{X}$ is a measurable space and $\phi : [0,\infty) \to (-\infty,\infty]$ is a convex function. $\mathbb{P} \ll \mathbb{Q}$ denotes that $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$. Well-known distance/divergence measures obtained by appropriately choosing ϕ include the Kullback-Leibler (KL) divergence ($\phi(t) = t\log t$), Hellinger distance ($\phi(t) = (\sqrt{t}-1)^2$), χ^2-divergence ($\phi(t) = (t-1)^2$) and total variation distance ($\phi(t) = |t-1|$).

Another popular family—particularly in probability theory and mathematical statistics—of distance measures on probabilities is the *integral probability metrics* (IPM) [55]—also called *probability metrics with a ζ-structure* [96]—defined as

$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{X}} f\,d\mathbb{P} - \int_{\mathcal{X}} f\,d\mathbb{Q} \right|, \tag{5.1}$$

where $\mathcal{F}$ in (5.1) is a class of real-valued bounded measurable functions on $\mathcal{X}$ (see Appendix A for the discussion on the relation between IPMs and ϕ-divergences wherein it is shown that IPMs are essentially different from ϕ-divergences). By appropriately choosing $\mathcal{F}$, various popular distance measures can be obtained:

(a) *Kantorovich metric, Wasserstein distance and Fortet-Mourier metric:* By setting $\mathcal{F} = \mathcal{F}_W := \{f : \|f\|_L \le 1\}$ in (5.1) yields the *Kantorovich metric*, W where $\|f\|_L$ is called the Lipschitz semi-norm of a bounded continuous real-valued function f on a metric space, $(\mathcal{X},\rho)$, with

$$\|f\|_L := \sup\left\{ \frac{|f(x)-f(y)|}{\rho(x,y)} \,:\, x \neq y \text{ in } \mathcal{X} \right\}.$$

The famous Kantorovich-Rubinstein theorem [23, Theorem 11.8.2] shows that when $\mathcal{X}$ is separable, the Kantorovich metric is the dual representation of *Wasserstein distance* [23, p. 420]—more specifically, the L^1-*Wasserstein distance*—defined as

$$W_1(\mathbb{P},\mathbb{Q}) := \inf_{\mu\in\mathcal{L}(\mathbb{P},\mathbb{Q})} \int \rho(x,y)\,d\mu(x,y), \tag{5.2}$$

where $\mathbb{P},\mathbb{Q} \in \{\mathbb{P} : \int \rho(x,y)\,d\mathbb{P}(x) < \infty, \forall\, y \in \mathcal{X}\}$ and $\mathcal{L}(\mathbb{P},\mathbb{Q})$ is the set of all measures on $\mathcal{X}\times\mathcal{X}$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. The L^1-Wasserstein distance (and therefore the Kantorovich metric) has found applications in information theory [35], mathematical statistics [59,96] and mass transportation problems [60].

The Fortet-Mourier metric [62, p. 17] is a generalization of the Kantorovich metric, with $\mathcal{F} := \{\|f\|_c \leq 1\}$, where

$$\|f\|_c := \sup\left\{\frac{|f(x) - f(y)|}{c(x,y)} : x \neq y \text{ in } \mathcal{X}\right\}$$

and $c(x,y) = \rho(x,y)\max(1, \rho(x,a)^{p-1}, \rho(y,a)^{p-1})$, $p \geq 1$ for some $a \in \mathcal{X}$. Note that when $p = 1$, the Fortet-Mourier metric is the same as the Kantorovich metric.

(b) *Dudley metric:* Choosing $\mathcal{F} = \mathcal{F}_\beta := \{f : \|f\|_{BL} \leq 1\}$ in (5.1) yields the *dual-bounded Lipschitz distance*—also called the *Dudley metric*, β—where

$$\|f\|_{BL} := \|f\|_\infty + \|f\|_L,$$

with $\|f\|_\infty := \sup\{|f(x)| : x \in \mathcal{X}\}$. The Dudley metric is popularly used in proving the convergence of probability measures with respect to the weak* (weak-star) topology on $M^1_+(\mathcal{X})$ [23, Chapter 11].

(c) *Total variation metric and Kolmogorov distance:* $\gamma_\mathcal{F}$ is the *total variation metric*, TV, when $\mathcal{F} = \mathcal{F}_{TV} := \{f : \|f\|_\infty \leq 1\}$ while it is the *Kolmogorov distance* when $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$. The Kolmogorov distance is popularly used in proving the classical central limit theorem in $\mathbb{R}^d$, and also appears as the Kolmogorov-Smirnov statistic in hypothesis testing [71].

(d) *Maximum mean discrepancy:* $\gamma_\mathcal{F}$ is called the *maximum mean discrepancy (MMD)* (see Proposition 3.2 and [37]) when $\mathcal{F} = \mathcal{F}_k := \{f : \|f\|_\mathcal{H} \leq 1\}$, where $\mathcal{H}$ represents an RKHS with a bounded and measurable reproducing kernel, k. MMD is used in statistical applications including homogeneity testing [37], independence testing [38], and testing for conditional independence [30].

Having mentioned various examples of IPMs and ϕ-divergences, we now consider the question of how is MMD related to other IPMs and ϕ-divergences, in particular its advantages and disadvantages over these families. This comparison is carried out on two fronts: (a) the ease of computation/estimation and (b) *strength* of the distance/divergence measure, which are elaborated below.

Computation and Estimation: The computation of ϕ-divergences and IPMs (including MMD) is not straightforward (for all $\mathbb{P}$ and $\mathbb{Q}$) as the integration in the former case and maximization in the latter case are not easily doable (for all $\mathbb{P}$ and $\mathbb{Q}$)—in the case of MMD, the integral is not easily computable for all $\mathbb{P}$, $\mathbb{Q}$ and k (see (3.5)). One approach to compute these distances between $\mathbb{P}$ and $\mathbb{Q}$ is to estimate them based on random samples drawn i.i.d. from them with the hope that the estimate converges to the true distance with large sample sizes. As aforementioned, this problem of estimating the distance between $\mathbb{P}$ and $\mathbb{Q}$ is also important in statistical inference applications (e.g., distribution testing) where $\mathbb{P}$ and $\mathbb{Q}$ are known only through random i.i.d. samples. For the estimators to be useful in practice, they have to: (i) be *consistent* (resp. *strongly consistent*), i.e., suppose $\{\theta_l\}$ is a sequence of estimators of θ, then θ_l is consistent (*resp.* strongly consistent) if θ_l converges in probability (*resp.* a.s.) to θ as $l \to \infty$, (ii) exhibit *fast* rate of convergence and (iii) be easy to implement. We use these properties to compare MMD to ϕ-divergences and other IPMs.

The non-parametric estimation of ϕ-divergence, especially the KL-divergence is a well-studied problem (see [56, 57, 89, 90] and references therein). Wang et al. [89] used a data-dependent space partitioning scheme and showed that the non-parametric estimator of KL-divergence is strongly consistent, while its rate of convergence can be arbitrarily slow depending on the distributions. In addition, for increasing dimensionality of the data (in $\mathcal{X} = \mathbb{R}^d$), the method is increasingly difficult to implement. On the other hand, by exploiting the variational representation of ϕ-divergences, Nguyen et al. [56, 57] provided a strongly consistent estimator of the KL-divergence by solving a convex quadratic program [10, Chapter 4]. Although this approach is efficient and the dimensionality of the data is not an issue, the rate of convergence of the estimator can be arbitrarily slow depending on the distributions.

Since we are not aware of any work on the non-parametric estimation of IPMs, in Section 5.2.1, we consider its estimation, in particular the Kantorovich metric (and therefore the L^1-Wasserstein distance), Dudley metric and MMD based on finite samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$. The *empirical estimators*—see (5.3)—

of the Kantorovich distance and Dudley metric are obtained by solving convex linear programs while that of MMD is computed in closed form, which means MMD is computationally simpler to estimate than that of ϕ-divergences and other IPMs.

Though various estimators may be used to estimate IPMs (and not just the empirical estimators), we show in Section 5.2.2 that the empirical estimators derived in Section 5.2.1 exhibit a nice connection to the problem of binary classification. In Section 5.2.2, we first show that $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ (*resp.* its empirical estimator) is the negative of the optimal risk associated with a binary classifier that separates the class conditional distributions, $\mathbb{P}$ and $\mathbb{Q}$ (*resp.* empirical distributions, $\mathbb{P}_m$ and $\mathbb{Q}_n$), where the classification rule is restricted to $\mathcal{F}$. In other words, the Kantorovich metric, Dudley metric and MMD (and their empirical estimators) can be understood as the negative of the optimal risk associated with a classifier for which the classification rule is restricted to $\mathcal{F}_W$, $\mathcal{F}_\beta$ and $\mathcal{F}_k$ respectively. We then show that the empirical estimators of the Kantorovich and Dudley metrics are related to the *margins* of the Lipschitz [88] and bounded Lipschitz classifiers, respectively; and MMD to the Parzen window classifier [68, 70] (see *kernel classification rule* [22, Chapter 10]) and support vector machine. The significance of this result is that the smoothness of the classifier is inversely related to the empirical estimator of the IPM between class conditionals $\mathbb{P}$ and $\mathbb{Q}$. Although this is intuitively clear, our result provides a theoretical justification.

Next, in Section 5.2.3, we show that the empirical estimators derived in Section 5.2.1 are strongly consistent and provide their rates of convergence, using concentration inequalities and tools from empirical process theory [86]. Based on these results, it will be clear that MMD exhibits fast rates of convergence compared to that of other IPMs (and ϕ-divergences) and its rate of convergence is independent of the dimension d (for $\mathcal{X} = \mathbb{R}^d$) unlike with other IPMs. Our experimental results in Section 5.2.4 confirm the convergence theory discussed in Section 5.2.3 and therefore demonstrate the practical viability of these estimators.

Since the total variation distance is also an IPM, in Section 5.2.5, we discuss its empirical estimator and show that it is not strongly consistent. Because of this, we provide new lower bounds for the total variation distance in terms of the Kantorovich metric, Dudley metric and MMD, which can be consistently estimated. These bounds also translate as lower bounds on the KL-divergence through Pinsker's inequality [25].

Strength of the Distance Measure: Let us consider the problem of estimating an unknown density based on finite random samples drawn i.i.d. from it. The quality of the estimate is measured by determining the distance between the estimated density and the true density. Given two probability metrics, ρ_1 and ρ_2, one might want to use the *stronger*[15] of the two to determine this distance, as the convergence of the estimated density to the true density in the stronger metric implies the convergence in the weaker metric, while the converse is not true. On the other hand, one might need to use a metric of weaker topology (i.e., coarser topology) to show convergence of some estimators, as the convergence might not occur w.r.t. a metric of strong topology. Since the relation between W, β, TV and KL-divergence is well-understood [33]—W and TV are stronger than β, KL-divergence is stronger than TV (by Pinsker's inequality), while no such relation exists between W and TV, though they are comparable when $\mathcal{X}$ is bounded—this motivates to study the relation between MMD and other IPMs to, e.g., determine which metrics are stronger respectively weaker. In Section 5.3, we show that MMD, i.e., γ_k is weaker than all these distance/divergence measures, wherein we just assume that k is measurable and bounded on $\mathcal{X}$. This means the topology induced by γ_k is coarser than the ones induced by all these metrics on $M^1_+(\mathcal{X})$. However, we show that if k is c_0-universal, then γ_k is equivalent to β, which is known to metrize the weak* (weak-star) topology (see Section 5.3 for details) on $M^1_+(\mathcal{X})$ [33, 71].

5.2 Empirical Estimation of IPMs

Given $\{X^{(1)}_1, X^{(1)}_2, \ldots, X^{(1)}_m\}$ and $\{X^{(2)}_1, X^{(2)}_2, \ldots, X^{(2)}_n\}$, which are i.i.d. samples drawn randomly from $\mathbb{P}$ and $\mathbb{Q}$ respectively, we propose to estimate $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$

[15] Two metrics $\rho_1, \rho_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ are said to be equivalent if $\rho_1(x, y) = 0 \Leftrightarrow \rho_2(x, y) = 0,\ x, y \in \mathcal{X}$. On the other hand, ρ_1 is said to be stronger than ρ_2 if $\rho_1(x, y) = 0 \Rightarrow \rho_2(x, y) = 0,\ x, y \in \mathcal{X}$ but not vice-versa. If ρ_1 is stronger than ρ_2, then we say ρ_2 is weaker than ρ_1.

by the following empirical estimator,

$$\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^{N} \widetilde{Y}_j f(X_j) \right|, \tag{5.3}$$

where $\mathbb{P}_m := \frac{1}{m}\sum_{j=1}^{m} \delta_{X_j^{(1)}}$ and $\mathbb{Q}_n := \frac{1}{n}\sum_{j=1}^{N} \delta_{X_j^{(2)}}$ represent the empirical distributions of $\mathbb{P}$ and $\mathbb{Q}$ respectively, $N = m + n$, $\widetilde{Y}_j = \frac{1}{m}$ when $X_j = X_j^{(1)}$ for $j = 1, \ldots, m$ and $\widetilde{Y}_{m+j} = -\frac{1}{n}$ when $X_{m+j} = X_j^{(2)}$ for $j = 1, \ldots, n$. Here, δ_x represents the Dirac measure at x. The computation of $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$ in (5.3) is not straightforward for any arbitrary $\mathcal{F}$. To obtain meaningful results, in Section 5.2.1, we restrict ourselves to $\mathcal{F}_W := \{f : \|f\|_L \leq 1\}$, $\mathcal{F}_\beta := \{f : \|f\|_{BL} \leq 1\}$ and $\mathcal{F}_k := \{f : \|f\|_{\mathcal{H}} \leq 1\}$ and compute (5.3), wherein we show that the Kantorovich (and therefore L^1-Wasserstein) and Dudley metrics can be estimated by solving linear programs (see Theorems 5.1 and 5.3) whereas an estimator for MMD can be obtained in closed form (Theorem 5.4; proved in [37]).

In Section 5.2.2, we present a novel interpretation of IPMs and their empirical estimators (especially of Kantorovich metric, Dudley metric and MMD) by relating them to binary classification.

5.2.1 Empirical Estimators of Kantorovich metric, Dudley metric and MMD

Let us denote $W := \gamma_{\mathcal{F}_W}$, $\beta := \gamma_{\mathcal{F}_\beta}$ and $\gamma_k := \gamma_{\mathcal{F}_k}$. The following results present the empirical estimators of Kantorovich metric (i.e., W), Dudley metric (i.e., β) and MMD (i.e., γ_k).

Theorem 5.1 (Empirical estimator of Kantorovich metric). *For all $\alpha \in [0, 1]$, the following function solves (5.3) for $\mathcal{F} = \mathcal{F}_W$:*

$$f_\alpha(x) := \alpha \min_{j=1,\ldots,N}(a_j^\star + \rho(x, X_j)) + (1-\alpha) \max_{j=1,\ldots,N}(a_j^\star - \rho(x, X_j)), \tag{5.4}$$

where

$$W(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{j=1}^{N} \widetilde{Y}_j a_j^\star, \tag{5.5}$$

and $\{a_j^\star\}_{j=1}^N$ *solve the following linear program,*

$$\max_{a_1,\ldots,a_N}\left\{\sum_{j=1}^N \widetilde{Y}_j a_j\,:\,-\rho(X_l,X_j)\le a_l-a_j\le\rho(X_l,X_j),\forall\,j,l\right\}. \tag{5.6}$$

Proof. Consider $W(\mathbb{P}_m,\mathbb{Q}_n)=\sup\{\sum_{j=1}^N\widetilde{Y}_jf(X_j):\|f\|_L\le 1\}$. Note that

$$1\ge\|f\|_L=\sup_{x\neq x'}\frac{|f(x)-f(x')|}{\rho(x,x')}\ge\max_{X_l\neq X_j}\frac{|f(X_l)-f(X_j)|}{\rho(X_l,X_j)},$$

which means

$$\begin{aligned}W(\mathbb{P}_m,\mathbb{Q}_n)&\le\sup\left\{\sum_{j=1}^N\widetilde{Y}_jf(X_j)\,:\,\max_{X_l\neq X_j}\frac{|f(X_l)-f(X_j)|}{\rho(X_l,X_j)}\le 1\right\}\\&=\sup\left\{\sum_{j=1}^N\widetilde{Y}_jf(X_j)\,:\,|f(X_l)-f(X_j)|\le\rho(X_l,X_j),\forall\,j,l\right\}\\&=\sup\left\{\sum_{j=1}^N\widetilde{Y}_ja_j\,:\,|a_l-a_j|\le\rho(X_l,X_j),\forall\,j,l\right\},\end{aligned}$$

where we have set $a_j:=f(X_j)$. Therefore, we have $W(\mathbb{P}_m,\mathbb{Q}_n)\le\sum_{j=1}^N\widetilde{Y}_ja_j^\star$, where $\{a_j^\star\}_{j=1}^N$ solve the linear program in (5.6). Note that the objective in (5.6) is linear in $\{a_j\}_{j=1}^N$ with linear inequality constraints and therefore by Theorem C.10, the optimum lies on the boundary of the constraint set, which means $\max_{X_l\neq X_j}\frac{|a_l^\star-a_j^\star|}{\rho(X_l,X_j)}=1$. Therefore, by Lemma C.4, f on $\{X_1,\ldots,X_N\}$ can be extended to a function f_α (on $\mathcal{X}$) defined in (5.4) where $f_\alpha(X_j)=f(X_j)=a_j^\star$ and $\|f_\alpha\|_L=\|f\|_L=1$, which means f_α is a maximizer of (5.3) and $W(\mathbb{P}_m,\mathbb{Q}_n)=\sum_{j=1}^N\widetilde{Y}_ja_j^\star$. □

Remark 5.2. *(a) The main result that is invoked in the proof of Theorem 5.1 is the extension of Lipschitz functions (defined on a subset of* $\mathcal{X}$*) to* $\mathcal{X}$*. Since such an extension is also possible for uniformly Hölder continuous functions, we obtain an empirical estimator of* $\gamma_{\mathcal{F}}$ *similar to (5.5) and (5.6)—with* ρ *in (5.6) replaced by* ρ^α*—where* $\mathcal{F}=\{\|f\|_\alpha\le 1\}$ *and*

$$\|f\|_\alpha:=\sup\left\{\frac{|f(x)-f(y)|}{\rho^\alpha(x,y)}\,:\,x\neq y\text{ in }\mathcal{X}\right\},\;0<\alpha\le 1.$$

(b) Applying the similar idea as in the proof of Theorem 5.1 to the empirical estimation of Fortet-Mourier metric, it can be shown that $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)\le\sum_{j=1}^N\widetilde{Y}_ja_j^\star$,

where $\{a_j^\star\}_{j=1}^N$ *solve the linear program in (5.6) with* $\rho(X_l, X_j)$ *replaced by* $c(X_l, X_j)$*. Since it is not clear whether an extension theorem similar to the one invoked in Theorem 5.1 (for Lipschitz functions) holds for* $f \in \{g : \|g\|_c < \infty\}$*, it is not clear whether* $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{j=1}^N \widetilde{Y}_j a_j^\star$ *holds for any* $\{X_j\}_{j=1}^N$.

Theorem 5.3 (Empirical estimator of Dudley metric)**.** *For all* $\alpha \in [0, 1]$*, the following function solves (5.3) for* $\mathcal{F} = \mathcal{F}_\beta$*:*

$$g_\alpha(x) := \max\left(-\max_{j=1,\ldots,N} |a_j^\star|, \min\left(h_\alpha(x), \max_{j=1,\ldots,N} |a_j^\star|\right)\right) \tag{5.7}$$

where

$$h_\alpha(x) := \alpha \min_{j=1,\ldots,N}(a_j^\star + L^\star \rho(x, X_j)) + (1-\alpha) \max_{j=1,\ldots,N}(a_j^\star - L^\star \rho(x, X_j)), \tag{5.8}$$

$$L^\star = \max_{X_l \neq X_j} \frac{|a_l^\star - a_j^\star|}{\rho(X_l, X_j)},$$

$$\beta(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{j=1}^N \widetilde{Y}_j a_j^\star, \tag{5.9}$$

and $\{a_j^\star\}_{j=1}^N$ *solve the following linear program,*

$$\begin{aligned} \max_{a_1,\ldots,a_N,b,c} \quad & \sum_{j=1}^N \widetilde{Y}_j a_j \\ s.t. \quad & -b\,\rho(X_l, X_j) \le a_l - a_j \le b\,\rho(X_l, X_j),\ \forall\, j, l \\ & -c \le a_j \le c,\ \forall\, j \\ & b + c \le 1. \end{aligned} \tag{5.10}$$

Proof. The proof is similar to that of Theorem 5.1. Note that

$$\begin{aligned} 1 \ge \|f\|_{BL} = \|f\|_L + \|f\|_\infty &= \sup_{x \neq y} \frac{|f(x) - f(y)|}{\rho(x, y)} + \sup_{x \in \mathcal{X}} |f(x)| \\ &\ge \max_{X_l \neq X_j} \frac{|f(X_l) - f(X_j)|}{\rho(X_l, X_j)} + \max_j |f(X_j)|, \end{aligned}$$

which means

$$\begin{aligned} \beta(\mathbb{P}_m, \mathbb{Q}_n) &= \sup\left\{\sum_{j=1}^N \widetilde{Y}_j f(X_j) \,:\, \|f\|_{BL} \le 1\right\} \\ &\le \sup\left\{\sum_{j=1}^N \widetilde{Y}_j f(X_j) : \max_j |f(X_j)| + \max_{X_l \neq X_j} \frac{|f(X_l) - f(X_j)|}{\rho(X_l, X_j)} \le 1\right\}. \end{aligned}$$

Let $a_j := f(X_j)$. Therefore, $\beta(\mathbb{P}_m, \mathbb{Q}_n) \le \sum_{j=1}^N \widetilde{Y}_j a_j^\star$, where $\{a_j^\star\}_{j=1}^N$ solve

$$\max_{a_1,\ldots,a_N} \left\{ \sum_{j=1}^N \widetilde{Y}_j a_j \,:\, \max_{X_l \neq X_j} \frac{|a_l - a_j|}{\rho(X_l, X_j)} + \max_j |a_j| \le 1 \right\}. \tag{5.11}$$

Introducing variables b and c such that $\max_{X_l \neq X_j} \frac{|a_l - a_j|}{\rho(X_l,X_j)} \le b$ and $\max_j |a_j| \le c$ reduces the program in (5.11) to (5.10). In addition, it is easy to see that the optimum occurs at the boundary of the constraint set, i.e., $\max_{X_l \neq X_j} \frac{|a_l - a_j|}{\rho(X_l,X_j)} + \max_j |a_j| = 1$. Hence, by Lemma C.5, g_α in (5.7) extends f defined on $\{X_1, \ldots, X_N\}$ to $\mathcal{X}$, i.e., $g_\alpha(X_j) = f(X_j)$, $\forall\, j$ and $\|g_\alpha\|_{BL} = \|f\|_{BL} = 1$. Note that h_α in (5.8) is the Lipschitz extension of f to $\mathcal{X}$ (by Lemma C.4). Therefore, g_α is a solution to (5.3) and (5.9) holds. □

Theorem 5.4 (Empirical estimator of MMD [37]). *For $\mathcal{F} = \mathcal{F}_k$, the following function is the unique solution to (5.3):*

$$f = \frac{1}{\|\sum_{j=1}^N \widetilde{Y}_j k(\cdot, X_j)\|_{\mathcal{H}}} \sum_{j=1}^N \widetilde{Y}_j k(\cdot, X_j), \tag{5.12}$$

and

$$\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) = \left\| \sum_{j=1}^N \widetilde{Y}_j k(\cdot, X_j) \right\|_{\mathcal{H}} = \sqrt{\sum_{l,j=1}^N \widetilde{Y}_l \widetilde{Y}_j k(X_l, X_j)}. \tag{5.13}$$

Proof. Consider $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) := \sup\{\sum_{j=1}^N \widetilde{Y}_j f(X_j) \,:\, \|f\|_{\mathcal{H}} \le 1\}$, which can be written as

$$\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) = \sup \left\{ \left\langle f, \sum_{j=1}^N \widetilde{Y}_j k(\cdot, X_j) \right\rangle_{\mathcal{H}} \,:\, \|f\|_{\mathcal{H}} \le 1 \right\},$$

where we have used the reproducing property of $\mathcal{H}$, i.e., $\forall\, f \in \mathcal{H}$, $\forall\, x \in \mathcal{X}$, $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$. The result follows from using the Cauchy-Schwartz inequality. □

It is clear from Theorems 5.1, 5.3 and 5.4 that the empirical estimator of MMD is very easy to implement (as it is available in a closed form) compared to those of Kantorovich and Dudley metrics, which involve solving linear programs. One important observation to be made about all these estimators is that they depend on $\{X_j\}_{j=1}^N$ through ρ or k, which means, once $\{\rho(X_j, X_j)\}_{i,j=1}^N$

or $\{k(X_l, X_j)\}_{j,l=1}^N$ is known, the complexity of the corresponding estimators is independent of the dimension d when $\mathcal{X} = \mathbb{R}^d$, unlike in the estimation of KL-divergence [89].

5.2.2 Interpretability of IPMs and their Empirical Estimators: Relation to Binary Classification

In this section, we provide a novel interpretation of IPMs and their empirical estimators by relating them to the problem of binary classification. We show in Proposition 5.5 that W, β and γ_k are the optimal risks associated with an appropriate binary classification problem, while in Proposition 5.6 we show their empirical estimators to be related to the margins of Lipschitz classifier [88], bounded Lipschitz classifier and support vector machine respectively. The significance of latter result is that the smoothness of these classifiers are inversely related to the distance between the empirical estimates of the class-conditional distributions, computed using W, β and γ_k respectively. In addition, we also establish the relation between MMD and the Parzen window classifier [68, 70] (also called the kernel classification rule [22, Chapter 10]).

Let us consider the binary classification problem with X being a $\mathcal{X}$-valued random variable, Y being a $\{-1, +1\}$-valued random variable and the product space, $\mathcal{X} \times \{-1, +1\}$, being endowed with a Borel probability measure μ. A discriminant function, f is a real valued measurable function on $\mathcal{X}$, whose sign is used to make a classification decision. Given a loss function, $L : \{-1, +1\} \times \mathbb{R} \to \mathbb{R}$, the goal is to choose an f that minimizes the risk associated with L, with the optimal L-risk being defined as,

$$\begin{aligned} R^L_{\mathcal{F}_\star} &= \inf_{f\in\mathcal{F}_\star} \int_{\mathcal{X}} L(y, f(x))\, d\mu(x,y) \\ &= \inf_{f\in\mathcal{F}_\star} \left\{ \varepsilon \int_{\mathcal{X}} L_1(f(x))\, d\mathbb{P}(x) + (1-\varepsilon) \int_{\mathcal{X}} L_{-1}(f(x))\, d\mathbb{Q}(x) \right\}, \end{aligned} \tag{5.14}$$

where $\mathcal{F}_\star$ is the set of all measurable functions on $\mathcal{X}$, $L_1(\alpha) := L(1, \alpha)$, $L_{-1}(\alpha) := L(-1, \alpha)$, $\mathbb{P}(X) := \mu(X|Y = +1)$, $\mathbb{Q}(X) := \mu(X|Y = -1)$, $\varepsilon := \mu(\mathcal{X}, Y = +1)$. Here, $\mathbb{P}$ and $\mathbb{Q}$ represent the class-conditional distributions and ε is the prior distribution of class +1. Now, we present the result that relates IPMs (between the class-conditional distributions) and the optimal L-risk of a binary classification problem.

Proposition 5.5 ($\gamma_\mathcal{F}$ and optimal L-risk)**.** *Let* $L_1(\alpha) = -\frac{\alpha}{\varepsilon}$ *and* $L_{-1}(\alpha) = \frac{\alpha}{1-\varepsilon}$. *Let* $\mathcal{F} \subset \mathcal{F}_\star$ *be such that* $f \in \mathcal{F} \Rightarrow -f \in \mathcal{F}$. *Then,* $\gamma_\mathcal{F}(\mathbb{P},\mathbb{Q}) = -R^L_\mathcal{F}$.

Proof. Define $\mathbb{P}f := \int_\mathcal{X} f\,d\mathbb{P}$. From (5.14), we have

$$\varepsilon \int_\mathcal{X} L_1(f)\,d\mathbb{P} + (1-\varepsilon)\int_\mathcal{X} L_{-1}(f)\,d\mathbb{Q} = \int_\mathcal{X} f\,d\mathbb{Q} - \int_\mathcal{X} f\,d\mathbb{P} = \mathbb{Q}f - \mathbb{P}f.$$

Therefore,

$$R^L_\mathcal{F} = \inf_{f\in\mathcal{F}}(\mathbb{Q}f - \mathbb{P}f) = -\sup_{f\in\mathcal{F}}(\mathbb{P}f - \mathbb{Q}f) \overset{(a)}{=} -\sup_{f\in\mathcal{F}}|\mathbb{P}f - \mathbb{Q}f| = -\gamma_\mathcal{F}(\mathbb{P},\mathbb{Q}),$$

where (a) follows from the fact that $\mathcal{F}$ is symmetric around zero, i.e., $f \in \mathcal{F} \Rightarrow -f \in \mathcal{F}$. □

Proposition 5.5 shows that $\gamma_\mathcal{F}(\mathbb{P},\mathbb{Q})$ (*resp.* $\gamma_\mathcal{F}(\mathbb{P}_m,\mathbb{Q}_n)$) is the negative of the optimal L-risk that is associated with a binary classifier that classifies the class-conditional distributions $\mathbb{P}$ and $\mathbb{Q}$ (*resp.* $\mathbb{P}_m$ and $\mathbb{Q}_n$) using the loss function, L, in Proposition 5.5, when the discriminant function is restricted to $\mathcal{F}$. Therefore, Theorem 5.5 provides a novel interpretation for the Kantorovich metric, Dudley metric and MMD (*resp.* their empirical estimators), which can be understood as the optimal L-risk associated with binary classifiers where the discriminant function, f is restricted to $\mathcal{F}_W$, $\mathcal{F}_\beta$ and $\mathcal{F}_k$ respectively. Proposition 5.5 also shows the importance of characteristic kernels in binary classification. This is because, if k is not characteristic, which means $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$ for some $\mathbb{P} \neq \mathbb{Q}$, then $R^L_{\mathcal{F}_k} = 0$, i.e., the risk is maximum (note that since $0 \le \gamma_k(\mathbb{P},\mathbb{Q}) = -R^L_{\mathcal{F}_k}$, the maximum risk is zero). In other words, if k is characteristic, then the maximum risk is obtained only when $\mathbb{P} = \mathbb{Q}$.

The following result (in Proposition 5.6) provides another interpretation for the empirical estimators of W, β and γ_k by relating them to the Lipschitz classifier, bounded Lipschitz classifier and support vector machine. Before we present the result, we briefly introduce these classifiers. Suppose $\{(X_j, Y_j)\}_{j=1}^N$, $X_j \in \mathcal{X}$, $Y_j \in$

$\{-1,+1\}, \forall\, j$ is a training sample drawn i.i.d. from μ and $m := |\{j : Y_j = +1\}|$. The Lipschitz classifier is defined as the solution, f_{lip} to the following program:

$$\inf\left\{\|f\|_L \,:\, f \in \text{Lip}(\mathcal{X},\rho),\, Y_j f(X_j) \geq 1,\, i = j,\ldots,N\right\}, \tag{5.15}$$

which is a large margin classifier with margin[16] $\frac{1}{\|f_{\text{lip}}\|_L}$. The program in (5.15) computes a *smooth* function, f that classifies the training sequence, $\{(X_j,Y_j)\}_{j=1}^N$ correctly (note that the constraints in (5.15) are such that $\text{sign}(f(X_j)) = Y_j$, which means f classifies the training sequence correctly, assuming the training sequence is separable). The smoothness is controlled by $\|f\|_L$ (the smaller the value of $\|f\|_L$, the smoother f and vice-versa). See [88] for a detailed study on the Lipschitz classifier. Replacing $\|f\|_L$ by $\|f\|_{BL}$ in (5.15) gives the bounded Lipschitz classifier, f_{BL} which is the solution to the following program:

$$\inf\left\{\|f\|_{BL} \,:\, f \in BL(\mathcal{X},\rho),\, Y_j f(X_j) \geq 1,\, j = 1,\ldots,N\right\}.$$

Note that replacing $\|f\|_L$ by $\|f\|_{\mathcal{H}}$ in (5.15), taking the infimum over $f \in \mathcal{H}$, yields the hard-margin support vector machine, f_{svm} [14], i.e.,

$$f_{\text{svm}} = \arg\inf\left\{\|f\|_{\mathcal{H}} \,:\, f \in \mathcal{H},\, Y_j f(X_j) \geq 1,\, j = 1,\ldots,N\right\}.$$

Proposition 5.6 (Empirical estimators and binary classification)**.** *The following hold:*

(a) $\frac{1}{\|f_{\text{lip}}\|_L} \leq \frac{1}{2}W(\mathbb{P}_m,\mathbb{Q}_n)$

(b) $\frac{1}{\|f_{\text{BL}}\|_{BL}} \leq \frac{1}{2}\beta(\mathbb{P}_m,\mathbb{Q}_n)$

(c) $\frac{1}{\|f_{\text{svm}}\|_{\mathcal{H}}} \leq \frac{1}{2}\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$*, if k is characteristic.*

To prove Proposition 5.6, we need the following lemma.

Lemma 5.7. *Let $\theta : V \to \mathbb{R}$ and $\psi : V \to \mathbb{R}$ be convex functions on a real vector space V. Suppose*

$$a = \sup\{\theta(x) : \psi(x) \leq b\}, \tag{5.16}$$

[16] The margin is a technical term used—in statistical machine learning—to indicate how well the training sample can be separated. Large margin classifiers (i.e., smooth classifiers) are preferred as they generalize well to unseen samples (i.e., test samples). See [68] for details.

where θ is not constant on $\{x : \psi(x) \le b\}$ and $a < \infty$. Then

$$b = \inf\{\psi(x) : \theta(x) \ge a\}. \tag{5.17}$$

Proof. Note that $A := \{x : \psi(x) \le b\}$ is a convex subset of V. Since θ is not constant on A, by Theorem C.10, θ attains its supremum on the boundary of A. Therefore, any solution, x_* to (5.16) satisfies $\theta(x_*) = a$ and $\psi(x_*) = b$. Let $G := \{x : \theta(x) > a\}$. For any $x \in G$, $\psi(x) > b$. If this were not the case, then x_* is not a solution to (5.16). Let $H := \{x : \theta(x) = a\}$. Clearly, $x_* \in H$ and so there exists an $x \in H$ for which $\psi(x) = b$. Suppose $\inf\{\psi(x) : x \in H\} = c < b$, which means for some $x^* \in H$, $x^* \in A$. From (5.16), this implies θ attains its supremum relative to A at some point of relative interior of A. By Theorem C.10, this implies θ is constant on A leading to a contradiction. Therefore, $\inf\{\psi(x) : x \in H\} = b$ and the result in (5.17) follows. □

Proof of Proposition 5.6. Define $\mathbb{P}f := \int_{\mathcal{X}} f\,d\mathbb{P}$. Note that $\|f\|_L$, $\|f\|_{BL}$ and $\|f\|_{\mathcal{H}}$ are convex functionals on the vector spaces $\text{Lip}(\mathcal{X},\rho)$, $BL(\mathcal{X},\rho)$ and $U(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{R} \,|\, \|f\|_{\mathcal{H}} < \infty\}$ respectively. Similarly, $\mathbb{P}f - \mathbb{Q}f$ is a convex functional on $\text{Lip}(\mathcal{X},\rho)$, $BL(\mathcal{X},\rho)$ and $U(\mathcal{X})$. Since $\mathbb{P} \ne \mathbb{Q}$, $\mathbb{P}f - \mathbb{Q}f$ is not constant on $\mathcal{F}_W$, $\mathcal{F}_\beta$ and $\mathcal{F}_k$. The results in (a)–(c) are obtained by appropriately choosing ψ, θ, V and b in Lemma 5.7. Here, we only prove (a) as the proofs of (b) and (c) are similar to that of (a).

Since $W(\mathbb{P}_m,\mathbb{Q}_n) = \sup\{\sum_{j=1}^N \widetilde{Y}_j f(X_j) \;:\; \|f\|_L \le 1\}$, by Lemma 5.7, we have

$$1 = \inf\left\{\|f\|_L : \sum_{j=1}^N \widetilde{Y}_j f(X_j) \ge W(\mathbb{P}_m,\mathbb{Q}_n),\ f \in \text{Lip}(\mathcal{X},\rho)\right\},$$

which can be written as

$$\frac{2}{W(\mathbb{P}_m,\mathbb{Q}_n)} = \inf\left\{\|f\|_L : \sum_{j=1}^N \widetilde{Y}_j f(X_j) \ge 2,\ f \in \text{Lip}(\mathcal{X},\rho)\right\}.$$

Note that $\{f \in \text{Lip}(\mathcal{X},\rho) : Y_j f(X_j) \ge 1, \forall\, j\} \subset \{f \in \text{Lip}(\mathcal{X},\rho) : \sum_{j=1}^N \widetilde{Y}_j f(X_j) \ge 2\}$, and therefore

$$\frac{2}{W(\mathbb{P}_m,\mathbb{Q}_n)} \le \inf\left\{\|f\|_L : Y_j f(X_j) \ge 1, \forall\, j,\ f \in \text{Lip}(\mathcal{X},\rho)\right\},$$

hence proving (a). Similar analysis for β and γ_k yield (b) and (c). □

The significance of this result is as follows. Proposition 5.6(a) shows that $\|f_{\text{lip}}\|_L \geq \frac{2}{W(\mathbb{P}_m,\mathbb{Q}_n)}$, which means the smoothness of the classifier, f_{lip}, computed as $\|f_{\text{lip}}\|_L$ is bounded by the inverse of the Kantorovich metric between $\mathbb{P}_m$ and $\mathbb{Q}_n$. So, if the distance between the class-conditionals $\mathbb{P}$ and $\mathbb{Q}$ is "small" (in terms of W), then the resulting Lipschitz classifier is less smooth, i.e., a "complex" classifier is required to classify the distributions $\mathbb{P}$ and $\mathbb{Q}$. A similar explanation holds for the bounded Lipschitz classifier and the support vector machine. Similar to Proposition 5.5, Proposition 5.6(c) also shows the importance of characteristic kernels in binary classification, especially in support vector machines. Suppose k is not characteristic, then $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ can be zero for $\mathbb{P}_m \neq \mathbb{Q}_n$, and therefore the margin is zero, which means even unlike distributions can become inseparable.

Based on Theorem 5.4 and Proposition 5.5, the empirical estimator of MMD can also be related to the Parzen window classifier as follows. Since the function, $f^* \in \mathscr{F}_k$ that achieves $R^L_{\mathscr{F}_k}$ (with L as in Proposition 5.5) is the same as the one in (5.12), the classification decision is given by

$$\text{sign}(f^*(x)) = \begin{cases} +1, & \frac{1}{m}\sum_{Y_j=1} k(x,X_j) > \frac{1}{n}\sum_{Y_j=-1} k(x,X_j) \\ -1, & \frac{1}{m}\sum_{Y_j=1} k(x,X_j) \leq \frac{1}{n}\sum_{Y_j=-1} k(x,X_j) \end{cases}, \tag{5.18}$$

which is exactly the classification function of a Parzen window classifier[17] [68, 70].

5.2.3 Consistency and Rate of Convergence

In Section 5.2.1, we presented the empirical estimators of W, β and γ_k. For these estimators to be reliable, we need them to converge to the population values as $m, n \to \infty$. Even if this holds, we would like to have a fast rate of convergence such that in practice, fewer samples are sufficient to obtain reliable estimates. We address these issues in this section. The strong consistency of $W(\mathbb{P}_m,\mathbb{Q}_n)$ and $\beta(\mathbb{P}_m,\mathbb{Q}_n)$ is shown in Proposition 5.9, while their rates of convergence are

[17] The classification rule in (5.18) differs from the "classical" Parzen window classifier in two respects. (i) Usually, the kernel (called the smoothing kernel) in the Parzen window rule is translation invariant in $\mathbb{R}^d$. In our case, $\mathcal{X}$ need not be $\mathbb{R}^d$ and k need not be translation invariant. So, the rule in (5.18) can be seen as a generalization of the classical Parzen window rule. (ii) The kernel in (5.18) is positive definite unlike in the classical Parzen window rule where k need not have to be so.

analyzed in Corollary 5.12. Corollary 5.12 also proves the strong consistency of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ and analyzes its rate of convergence. We show that $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ enjoys a fast rate of convergence compared to $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$.

Before we start presenting the results, we briefly introduce some terminology and notation from empirical process theory. For any $r \geq 1$ and probability measure $\mathbb{Q}$, define the L^r norm $\|f\|_{\mathbb{Q},r} := (\int_{\mathcal{X}} |f|^r \, d\mathbb{Q})^{1/r}$ and let $L^r(\mathbb{Q})$ denote the metric space induced by this norm. The *covering number* $\mathcal{N}(\mathcal{F}, L^r(\mathbb{Q}), \varepsilon)$ is the minimal number of $L^r(\mathbb{Q})$ balls of radius ε needed to cover $\mathcal{F}$. $\mathcal{H}(\mathcal{F}, L^r(\mathbb{Q}), \varepsilon) := \log \mathcal{N}(\mathcal{F}, L^r(\mathbb{Q}), \varepsilon)$ is called the *entropy* of $\mathcal{F}$ using the $L^r(\mathbb{Q})$ metric. Define the minimal envelope function: $F(x) := \sup_{f \in \mathcal{F}} |f(x)|$.

We now present a general result on the strong consistency of $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$, which simply follows from Theorem C.11.

Lemma 5.8. *Suppose the following conditions hold:*

(i) $\int F \, d\mathbb{P} < \infty$.

(ii) $\int F \, d\mathbb{Q} < \infty$.

(iii) $\forall \varepsilon > 0$, $\frac{1}{m}\mathcal{H}(\mathcal{F}, L^1(\mathbb{P}_m), \varepsilon) \xrightarrow{\mathbb{P}} 0$ *as* $m \to \infty$.

(iv) $\forall \varepsilon > 0$, $\frac{1}{n}\mathcal{H}(\mathcal{F}, L^1(\mathbb{Q}_n), \varepsilon) \xrightarrow{\mathbb{Q}} 0$ *as* $n \to \infty$.

Then, $|\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})| \xrightarrow{a.s.} 0$ *as* $m, n \to \infty$.

Proof. Define $\mathbb{P}f := \int_{\mathcal{X}} f \, d\mathbb{P}$. Note that $|\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})| \leq \sup_{f \in \mathcal{F}} |\mathbb{P}_m f - \mathbb{P}f| + \sup_{f \in \mathcal{F}} |\mathbb{Q}_n f - \mathbb{Q}f|$. Therefore, by Theorem C.11, $\sup_{f \in \mathcal{F}} |\mathbb{P}_m f - \mathbb{P}f| \xrightarrow{a.s.} 0$, $\sup_{f \in \mathcal{F}} |\mathbb{Q}_n f - \mathbb{Q}f| \xrightarrow{a.s.} 0$ and the result follows. □

The following corollary to Lemma 5.8 shows that $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ are strongly consistent.

Proposition 5.9 (Consistency of W and β)**.** *Let* $(\mathcal{X}, \rho)$ *be a totally bounded metric space. Then, as* $m, n \to \infty$,

(i) $|W(\mathbb{P}_m, \mathbb{Q}_n) - W(\mathbb{P}, \mathbb{Q})| \xrightarrow{a.s.} 0$.

(ii) $|\beta(\mathbb{P}_m, \mathbb{Q}_n) - \beta(\mathbb{P}, \mathbb{Q})| \xrightarrow{a.s.} 0$.

Proof. For any $f \in \mathcal{F}_W$,

$$f(x) \le \sup_{x\in\mathcal{X}} |f(x)| \le \sup_{x,y} |f(x) - f(y)| \le \|f\|_L \sup_{x,y} \rho(x,y) \le \|f\|_L \text{diam}(\mathcal{X}) < \infty,$$

where $\text{diam}(\mathcal{X})$ represents the diameter of $\mathcal{X}$. Therefore, $\forall\, x \in \mathcal{X}$, $F(x) \le \text{diam}(\mathcal{X}) < \infty$, which satisfies (i) and (ii) in Lemma 5.8. Here $\text{diam}(\mathcal{X}) := \sup\{\rho(x,y) : x, y \in \mathcal{X}\}$. Kolmogorov and Tihomirov [45] have shown that

$$\mathcal{H}(\mathcal{F}_W, \|\cdot\|_\infty, \varepsilon) \le \mathcal{N}\left(\mathcal{X}, \rho, \frac{\varepsilon}{4}\right) \log\left(2\left\lceil \frac{2\,\text{diam}(\mathcal{X})}{\varepsilon}\right\rceil + 1\right). \tag{5.19}$$

Since $\mathcal{H}(\mathcal{F}_W, L^1(\mathbb{P}_m), \varepsilon) \le \mathcal{H}(\mathcal{F}_W, \|\cdot\|_\infty, \varepsilon)$, the conditions (iii) and (iv) in Lemma 5.8 are satisfied and therefore, $|W(\mathbb{P}_m, \mathbb{Q}_n) - W(\mathbb{P}, \mathbb{Q})| \xrightarrow{a.s.} 0$ as $m, n \to \infty$. Since $\mathcal{F}_\beta \subset \mathcal{F}_W$, the envelope function associated with $\mathcal{F}_\beta$ is upper bounded by the envelope function associated with $\mathcal{F}_W$ and $\mathcal{H}(\mathcal{F}_\beta, \|\cdot\|_\infty, \varepsilon) \le \mathcal{H}(\mathcal{F}_W, \|\cdot\|_\infty, \varepsilon)$. Therefore, the result for β follows. □

Similar to Proposition 5.9, a strong consistency result for γ_k can be provided by estimating the entropy number of $\mathcal{F}_k$. See Cucker and Zhou [16, Chapter 5] for the estimates of entropy numbers for various $\mathcal{H}$. However, in the following, we adopt a different approach to prove the strong consistency of γ_k. To this end, we first provide a general result on the rate of convergence of $\gamma_\mathcal{F}(\mathbb{P}_m, \mathbb{Q}_n)$ and then, as a special case, obtain the rates of convergence of the empirical estimators of W, β and γ_k. Using this result, we then prove the strong consistency of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$. We start with the following definition.

Definition 5.10 (Rademacher complexity)**.** *Let $\mathcal{F}$ be a class of functions on $\mathcal{X}$ and $\{\varrho_j\}_{j=1}^m$ be independent Rademacher random variables, i.e., $Pr(\varrho_j = +1) = Pr(\varrho_j = -1) = \frac{1}{2}$. The Rademacher process is defined as $\{\frac{1}{m}\sum_{j=1}^m \varrho_i f(X_j) : f \in \mathcal{F}\}$ for some $\{X_j\}_{j=1}^m \subset \mathcal{X}$. The Rademacher complexity over $\mathcal{F}$ is defined as*

$$R_m(\mathcal{F}; \{X_j\}_{j=1}^m) := \mathbb{E} \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{j=1}^m \varrho_j f(X_j) \right|.$$

We now present a general result that provides a probabilistic bound on the deviation of $\gamma_\mathcal{F}(\mathbb{P}_m, \mathbb{Q}_n)$ from $\gamma_\mathcal{F}(\mathbb{P}, \mathbb{Q})$. This generalizes Theorem 4 in [37], the main difference being that we now consider function classes other than RKHSs, and thus express the bound in terms of the Rademacher complexities (see the proof for further discussion).

Theorem 5.11. *For any $\mathcal{F}$ such that $\nu := \sup_{x\in\mathcal{X}} F(x) < \infty$, with probability at least $1-\delta$, the following holds:*

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| \le \sqrt{18\nu^2\log\frac{4}{\delta}}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)+2R_m(\mathcal{F};\{X_j^{(1)}\}_{j=1}^m) + 2R_n(\mathcal{F};\{X_j^{(2)}\}_{j=1}^n). \tag{5.20}$$

Proof. Define $\mathbb{P}f := \int_{\mathcal{X}} f\,d\mathbb{P}$. Since $|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| \le \sup_{f\in\mathcal{F}}|\mathbb{P}_mf-\mathbb{P}f|+\sup_{f\in\mathcal{F}}|\mathbb{Q}_nf-\mathbb{Q}f|$, we bound the terms $\sup_{f\in\mathcal{F}}|\mathbb{P}_mf-\mathbb{P}f|$ and $\sup_{f\in\mathcal{F}}|\mathbb{Q}_nf-\mathbb{Q}f|$, which are the fundamental quantities that appear in empirical process theory. The proof strategy begins in a manner similar to [36, Appendix A.2], but with an additional step which will be flagged below.

Note that $\sup_{f\in\mathcal{F}}|\mathbb{P}_mf-\mathbb{P}f|$ satisfies (C.6) with $c_i=\frac{2\nu}{m}$. Therefore, by McDiarmid's inequality in (C.7), we have that with probability at least $1-\frac{\delta}{4}$, the following holds:

$$\begin{aligned}\sup_{f\in\mathcal{F}}|\mathbb{P}_mf-\mathbb{P}f| &\le \mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{P}_mf-\mathbb{P}f|+\sqrt{\frac{2\nu^2}{m}\log\frac{4}{\delta}}\\ &\overset{(a)}{\le} 2\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{j=1}^m\varrho_jf(X_j^{(1)})\right|+\sqrt{\frac{2\nu^2}{m}\log\frac{4}{\delta}},\\ &= 2\,\mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{j=1}^m\varrho_jf(X_j^{(1)})\right|\,\middle|\,\{X_j^{(1)}\}_{j=1}^m\right]\right]+\sqrt{\frac{2\nu^2}{m}\log\frac{4}{\delta}},\end{aligned} \tag{5.21}$$

where (a) follows from bounding $\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{P}_mf-\mathbb{P}f|$ by using the symmetrization inequality in (C.8).

Since $\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\frac{1}{m}\sum_{j=1}^m\varrho_jf(X_j^{(1)})|\mid\{X_j^{(1)}\}_{j=1}^m\right]$ satisfies (C.6) with $c_i=\frac{2\nu}{m}$, by McDiarmid's inequality in (C.7), with probability at least $1-\frac{\delta}{4}$, we have

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{j=1}^m\varrho_jf(X_j^{(1)})\right| \le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{j=1}^m\varrho_jf(X_j^{(1)})\right|\,\middle|\,\{X_j^{(1)}\}_{j=1}^m\right]+\sqrt{\frac{2\nu^2}{m}\log\frac{4}{\delta}}. \tag{5.22}$$

Tying (5.21) and (5.22), we have that with probability at least $1-\frac{\delta}{2}$, the following holds:

$$\sup_{f\in\mathcal{F}}|\mathbb{P}_m f-\mathbb{P}f|\le 2R_m(\mathcal{F};\{X_i^{(1)}\}_{j=1}^m)+\sqrt{\frac{18\nu^2}{m}\log\frac{4}{\delta}}. \tag{5.23}$$

Performing similar analysis for $\sup_{f\in\mathcal{F}}|\mathbb{Q}_n f-\mathbb{Q}f|$, we have that with probability at least $1-\frac{\delta}{2}$,

$$\sup_{f\in\mathcal{F}}|\mathbb{Q}_n f-\mathbb{Q}f|\le 2R_n(\mathcal{F};\{X_j^{(2)}\}_{j=1}^n)+\sqrt{\frac{18\nu^2}{n}\log\frac{4}{\delta}}. \tag{5.24}$$

The result follows by adding (5.23) and (5.24). Note that the second application of McDiarmid was not needed in [36, Appendix A.2], since in that case a simplification was possible due to $\mathcal{F}$ being restricted to RKHSs. □

Theorem 5.11 holds for any $\mathcal{F}$ for which ν is finite. However, to obtain the rate of convergence for $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)$, one requires an estimate of $R_m(\mathcal{F};\{X_j^{(1)}\}_{j=1}^m)$ and $R_n(\mathcal{F};\{X_j^{(2)}\}_{j=1}^n)$. Note that if $R_m(\mathcal{F};\{X_j^{(1)}\}_{j=1}^m)\xrightarrow{\mathbb{P}}0$ as $m\to\infty$ and $R_n(\mathcal{F};\{X_j^{(2)}\}_{j=1}^n)\xrightarrow{\mathbb{Q}}0$ as $n\to\infty$, then

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})|\xrightarrow{\mathbb{P},\mathbb{Q}}0 \text{ as } m,n\to\infty.$$

Also note that if $R_m(\mathcal{F};\{X_j^{(1)}\}_{j=1}^m)=O_{\mathbb{P}}(r_m)$ and $R_n(\mathcal{F};\{X_j^{(2)}\}_{j=1}^n)=O_{\mathbb{Q}}(r_n)$ where $r_m,r_n\to 0$ as $m,n\to\infty$, then from (5.20),

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})|=O_{\mathbb{P},\mathbb{Q}}(r_m\vee m^{-1/2}+r_n\vee n^{-1/2}),$$

where $a\vee b:=\max(a,b)$. The following corollary to Theorem 5.11 provides the rate of convergence for $W(\mathbb{P}_m,\mathbb{Q}_n)$, $\beta(\mathbb{P}_m,\mathbb{Q}_n)$ and $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$. Note that Corollary 5.12(ii) was proved in [37], [36, Appendix A.2] by a more direct argument, where the fact that $\mathcal{F}_k$ is an RKHS was used at an earlier stage of the proof to simplify the reasoning. We include the result here for completeness.

Corollary 5.12 (Rates of convergence for W, β and γ_k)**.** (i) *Let $\mathcal{X}$ be a bounded subset of $(\mathbb{R}^d,\|\cdot\|_s)$ for some $1\le s\le\infty$. Then,*

$$|W(\mathbb{P}_m,\mathbb{Q}_n)-W(\mathbb{P},\mathbb{Q})|=O_{\mathbb{P},\mathbb{Q}}(r_m+r_n)$$

and

$$|\beta(\mathbb{P}_m,\mathbb{Q}_n)-\beta(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_m+r_n),$$

where

$$r_m = \begin{cases} m^{-1/2}\log m, & d=1 \\ m^{-1/(d+1)}, & d\geq 2 \end{cases}. \tag{5.25}$$

In addition if $\mathcal{X}$ is a bounded, convex subset of $(\mathbb{R}^d, \|\cdot\|_s)$ with non-empty interior, then

$$r_m = \begin{cases} m^{-1/2}, & d=1 \\ m^{-1/2}\log m, & d=2 \\ m^{-1/d}, & d>2 \end{cases}. \tag{5.26}$$

(ii) *Let $\mathcal{X}$ be a measurable space. Suppose k is measurable and $\sup_{x\in\mathcal{X}} k(x,x) \leq C < \infty$. Then,*

$$|\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_k(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(m^{-1/2}+n^{-1/2}).$$

In addition,

$$|\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_k(\mathbb{P},\mathbb{Q})| \xrightarrow{a.s.} 0 \ \text{ as } \ m,n\to\infty,$$

i.e., the empirical estimator of MMD is strongly consistent.

Proof. (i) The generalized entropy bound [88, Theorem 16] gives that for every $\varepsilon > 0$,

$$R_m(\mathcal{F}; \{X_j^{(1)}\}_{j=1}^m) \leq 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{m}}\int_{\varepsilon/4}^{\infty}\sqrt{\mathcal{H}(\mathcal{F}, L^2(\mathbb{P}_m),\tau)}\,d\tau. \tag{5.27}$$

Let $\mathcal{F}=\mathcal{F}_W$. Since $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$, it is totally bounded and therefore the entropy number in (5.27) can be bounded through (5.19) by noting that

$$\mathcal{H}(\mathcal{F}_W, L^2(\mathbb{P}_m),\tau) \leq \mathcal{H}(\mathcal{F}_W, \|\cdot\|_\infty,\tau) \leq \frac{C_1}{\tau^{d+1}} + \frac{C_2}{\tau^d}, \tag{5.28}$$

where we have used the fact that $\mathcal{N}(\mathcal{X},\|\cdot\|_s,\varepsilon) = O(\varepsilon^{-d})$, $1\leq s\leq\infty$ and $\log(\lceil x\rceil+1)\leq x+1$.[18] The constants C_1 and C_2 depend only on the properties of $\mathcal{X}$ and are independent of τ. Substituting (5.28) in (5.27), we have

$$R_m(\mathcal{F}_W; \{X_j^{(1)}\}_{j=1}^m) \leq \inf_{\varepsilon>0}\left[2\varepsilon + \frac{4\sqrt{2}}{\sqrt{m}}\int_{\varepsilon/4}^{\infty}\sqrt{\mathcal{H}(\mathcal{F}_W, L^2(\mathbb{P}_m),\tau)}\,d\tau\right]$$

[18] Note that for any $x\in\mathcal{X}\subset\mathbb{R}^d$, $\|x\|_\infty\leq\cdots\leq\|x\|_s\leq\cdots\leq\|x\|_2\leq\|x\|_1\leq\sqrt{d}\|x\|_2$. Therefore, $\forall\, s\geq 2$, $\mathcal{N}(\mathcal{X},\|\cdot\|_s,\varepsilon)\leq\mathcal{N}(\mathcal{X},\|\cdot\|_2,\varepsilon)$ and $\forall\, 1\leq s\leq 2$, $\mathcal{N}(\mathcal{X},\|\cdot\|_s,\varepsilon)\leq\mathcal{N}(\mathcal{X},\sqrt{d}\|\cdot\|_2,\varepsilon)=\mathcal{N}(\mathcal{X},\|\cdot\|_2,\varepsilon/\sqrt{d})$. Use $\mathcal{N}(\mathcal{X},\|\cdot\|_2,\varepsilon)=O(\varepsilon^{-d})$ [85, Lemma 2.5].

$$\leq \inf_{\varepsilon>0}\left[2\varepsilon+\frac{4\sqrt{2}}{\sqrt{m}}\int_{\varepsilon/4}^{4R}\left(\frac{\sqrt{C_1}}{\tau^{(d+1)/2}}+\frac{\sqrt{C_2}}{\tau^{d/2}}\right)\,d\tau\right],$$

where $R := \mathrm{diam}(\mathcal{X})$. Note the change in upper limits of the integral from ∞ to $4R$. This is because $\mathcal{X}$ is totally bounded and $\mathcal{H}(\mathcal{F}_W, \|\cdot\|_\infty, \tau)$ depends on $\mathcal{N}(\mathcal{X}, \|\cdot\|_s, \tau/4)$. The rates in (5.25) are simply obtained by solving the right hand side of the above inequality. As mentioned in the paragraph preceding the statement of Corollary 5.12, we have $r_m \vee m^{-1/2} = r_m$ and so the result for $W(\mathbb{P}_m, \mathbb{Q}_n)$ follows.

Suppose $\mathcal{X}$ is convex. Then $\mathcal{X}$ is connected. It is easy to see that $\mathcal{X}$ is also centered, i.e., for all subsets $A \subset \mathcal{X}$ with $\mathrm{diam}(A) \leq 2r$ there exists a point $x \in \mathcal{X}$ such that $\|x-a\|_s \leq r$ for all $a \in A$. Since $\mathcal{X}$ is connected and centered, we have from [45] that

$$\begin{aligned}\mathcal{H}(\mathcal{F}_W, L^2(\mathbb{P}_m), \tau) &\leq \mathcal{H}(\mathcal{F}_W, \|\cdot\|_\infty, \tau)\\ &\leq \mathcal{N}\left(\mathcal{X}, \|\cdot\|_s, \frac{\tau}{2}\right)\log 2 + \log\left(2\left\lceil\frac{2\,\mathrm{diam}(\mathcal{X})}{\tau}\right\rceil + 1\right)\\ &\leq C_3\tau^{-d} + C_4\tau^{-1} + C_5, \end{aligned} \tag{5.29}$$

where we used the fact that $\mathcal{N}(\mathcal{X}, \|\cdot\|_s, \varepsilon) = O(\varepsilon^{-d})$. C_3, C_4 and C_5 are constants that depend only on the properties of $\mathcal{X}$ and are independent of τ. Substituting (5.29) in (5.27), we have,

$$R_m(\mathcal{F}_W; \{X_j^{(1)}\}_{j=1}^m) \leq \inf_{\varepsilon>0}\left[2\varepsilon + \frac{4\sqrt{2}}{\sqrt{m}}\int_{\varepsilon/4}^{2R}\frac{\sqrt{C_3}}{\tau^{d/2}}\,d\tau\right] + O(m^{-1/2}).$$

Again note the change in upper limits of the integral from ∞ to $2R$. This is because $\mathcal{H}(\mathcal{F}_W, \|\cdot\|_\infty, \tau)$ depends on $\mathcal{N}(\mathcal{X}, \|\cdot\|_s, \tau/2)$. The rates in (5.26) are obtained by solving the right hand side of the above inequality. Since $r_m \vee m^{-1/2} = r_m$, the result for $W(\mathbb{P}_m, \mathbb{Q}_n)$ follows.

Since $\mathcal{F}_\beta \subset \mathcal{F}_W$, we have $R_m(\mathcal{F}_\beta; \{X_j^{(1)}\}_{j=1}^m) \leq R_m(\mathcal{F}_W; \{X_j^{(1)}\}_{j=1}^m)$ and therefore, the result for $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ follows. The rates in (5.26) can also be directly obtained for β by using the entropy number of $\mathcal{F}_\beta$, i.e., $\mathcal{H}(\mathcal{F}_\beta, \|\cdot\|_\infty, \varepsilon) = O(\varepsilon^{-d})$ [86, Theorem 2.7.1] in (5.27).

(ii) By [6, Lemma 22], $R_m(\mathcal{F}_k; \{X_j^{(1)}\}_{j=1}^m) \leq \frac{\sqrt{C}}{\sqrt{m}}$ and $R_n(\mathcal{F}_k; \{X_j^{(2)}\}_{j=1}^n) \leq \frac{\sqrt{C}}{\sqrt{n}}$. Substituting these in (5.20) yields the result. In addition, by the Borel-Cantelli lemma [23, Theorem 8.3.4], the strong consistency of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ follows. □

Remark 5.13. *(i) Note that the rate of convergence of $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ is dependent on the dimension, d, which means that in large dimensions, more samples are needed to obtain useful estimates of $W(\mathbb{P}, \mathbb{Q})$ and $\beta(\mathbb{P}, \mathbb{Q})$. Also note that the rates are independent of the metric, $\|\cdot\|_s$, $1 \le s \le \infty$.*

(ii) When $\mathcal{X}$ is a bounded, convex subset of $(\mathbb{R}^d, \|\cdot\|_s)$, faster rates are obtained than for the case where $\mathcal{X}$ is just a bounded (but not convex) subset of $(\mathbb{R}^d, \|\cdot\|_s)$.

(iii) In the case of MMD, we have not made any assumptions on $\mathcal{X}$ except it being a topological space. When $\mathcal{X} = \mathbb{R}^d$, the rate is independent of d, which is a very useful property. The condition of the kernel being bounded is satisfied by a host of kernels, the examples of which include the Gaussian kernel, Laplacian kernel, inverse multiquadratics, etc., on $\mathbb{R}^d$. See Wendland [91] for more examples. As mentioned before, the estimates for $R_m(\mathcal{F}_k; \{X_j^{(1)}\}_{j=1}^m)$ can be directly obtained by using the entropy numbers of $\mathcal{F}_k$. See Cucker and Zhou [16, Chapter 5] for the estimates of entropy numbers for various $\mathcal{H}$.

To summarize, in this section, we have shown that the empirical estimators of Kantorovich metric, Dudley metric and MMD are strongly consistent and the empirical estimator of MMD exhibits fast rate of convergence compared to those of Kantorovich and Dudley metrics (and also of ϕ-divergence [56, 57, 89]). Therefore, based on the results in this section and Section 5.2.1, it is clear that the empirical estimator of MMD has more favorable properties compared to the other empirical estimators under consideration and hence is more suited for use in statistical inference applications like two-sample tests [37].

5.2.4 Simulation Results

So far, in Sections 5.2.1 and 5.2.3, we have presented the empirical estimators of W, β and γ_k and their convergence analysis. Though we have shown that the empirical estimator of γ_k has more favorable properties than those of W and β, we would like to know how good these estimators are in practice? In this section, we demonstrate the performance of these estimators through simulations.

As we have mentioned before, given $\mathbb{P}$ and $\mathbb{Q}$, it is usually difficult to exactly compute $W(\mathbb{P},\mathbb{Q})$, $\beta(\mathbb{P},\mathbb{Q})$ and $\gamma_k(\mathbb{P},\mathbb{Q})$. However, in order to test the performance of their empirical estimators, in the following, we consider some examples where $W(\mathbb{P},\mathbb{Q})$, $\beta(\mathbb{P},\mathbb{Q})$ and $\gamma_k(\mathbb{P},\mathbb{Q})$ can be computed exactly.

Empirical Estimator of $W(\mathbb{P},\mathbb{Q})$

For the ease of computation, let us consider $\mathbb{P}$ and $\mathbb{Q}$ (defined on the Borel σ-algebra of $\mathbb{R}^d$) as product measures, $\mathbb{P} = \otimes_{j=1}^d \mathbb{P}^{(j)}$ and $\mathbb{Q} = \otimes_{j=1}^d \mathbb{Q}^{(j)}$, where $\mathbb{P}^{(j)}$ and $\mathbb{Q}^{(j)}$ are defined on the Borel σ-algebra of $\mathbb{R}$. In this setting, when $\rho(x,y) = \|x-y\|_1$, it is easy to show that

$$W(\mathbb{P},\mathbb{Q}) = \sum_{j=1}^d W(\mathbb{P}^{(j)},\mathbb{Q}^{(j)}), \tag{5.30}$$

where

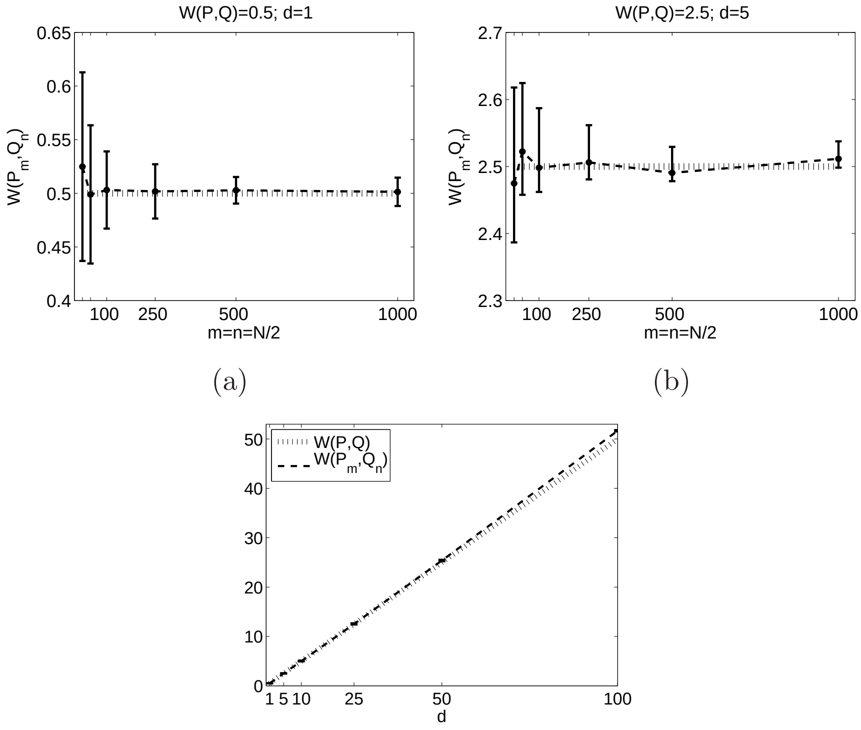
$$W(\mathbb{P}^{(j)},\mathbb{Q}^{(j)}) = \int_{\mathbb{R}} \left|F_{\mathbb{P}^{(j)}}(x) - F_{\mathbb{Q}^{(j)}}(x)\right| \, dx, \tag{5.31}$$

and $F_{\mathbb{P}^{(j)}}(x) = \mathbb{P}^{(j)}((-\infty,x])$ [84].[19] Now, in the following, we consider two examples where W in (5.31) can be computed in closed form. Note that we need $\mathcal{X}$ to be a bounded subset of $\mathbb{R}^d$ such that the consistency of $W(\mathbb{P}_m,\mathbb{Q}_n)$ is guaranteed by Corollary 5.12.

Example 5.14. *Let $\mathcal{X} = \times_{j=1}^d [a_j, s_j]$. Suppose $\mathbb{P}^{(j)} = U[a_j,b_j]$ and $\mathbb{Q}^{(j)} = U[r_j,s_j]$, which are uniform distributions on $[a_j,b_j]$ and $[r_j,s_j]$ respectively, where $-\infty < a_j \le r_j \le b_j \le s_j < \infty$. Then, it is easy to verify that $W(\mathbb{P}^{(j)},\mathbb{Q}^{(j)}) = (s_j + r_j - a_j - b_j)/2$ and $W(\mathbb{P},\mathbb{Q})$ follows from (5.30).*

Figures 5.1(a) and 5.1(b) show $W(\mathbb{P}_m,\mathbb{Q}_n)$ (shown in thick dotted lines) for $d = 1$ and $d = 5$ respectively. Figure 5.1(c) shows the behavior of $W(\mathbb{P}_m,\mathbb{Q}_n)$ and $W(\mathbb{P},\mathbb{Q})$ for various d with a fixed sample size of $m = n = 250$. Here, we chose $a_j = -\frac{1}{2}$, $b_j = \frac{1}{2}$, $r_j = 0$ and $s_j = 1$ for all $j = 1,\ldots,d$ such that $W(\mathbb{P}^{(j)},\mathbb{Q}^{(j)}) = \frac{1}{2}$, $\forall j$ and $W(\mathbb{P},\mathbb{Q}) = \frac{d}{2}$, shown in thin dotted lines in Figures 5.1(a-c).

[19] The explicit form for the L^1-Wasserstein distance in (5.2) is known for $(\mathcal{X},\rho(x,y)) = (\mathbb{R},|x-y|)$ [83, 84], which is given as $W_1(\mathbb{P},\mathbb{Q}) = \int_{(0,1)} |F_{\mathbb{P}}^{-1}(u) - F_{\mathbb{Q}}^{-1}(u)|\,du = \int_{\mathbb{R}} |F_{\mathbb{P}}(x) - F_{\mathbb{Q}}(x)|\,dx$, where $F_{\mathbb{P}}(x) = \mathbb{P}((-\infty,x])$ and $F_{\mathbb{P}}^{-1}(u) = \inf\{x \in \mathbb{R} | F_{\mathbb{P}}(x) \ge u\}$, $0 < u < 1$. However, the exact computation (in closed form) of $W_1(\mathbb{P},\mathbb{Q})$ is not straightforward for all $\mathbb{P}$ and $\mathbb{Q}$. Note that since $\mathbb{R}^d$ is separable, by the Kantorovich-Rubinstein theorem, $W(\mathbb{P},\mathbb{Q}) = W_1(\mathbb{P},\mathbb{Q})$, $\forall\,\mathbb{P},\mathbb{Q}$.



(a)

(b)

(c)

Figure 5.1: (a-b) represent the empirical estimates of the Kantorovich metric (shown in thick dotted lines) between $\mathbb{P} = U[-\frac{1}{2}, \frac{1}{2}]^d$ and $\mathbb{Q} = U[0,1]^d$ with $\rho(x,y) = \|x-y\|_1$, for increasing sample size N, where $d=1$ in (a) and $d=5$ in (b). Here $U[l_1, l_2]^d$ represents a uniform distribution on $[l_1, l_2]^d$ (see Example 5.14 for details). The population values of the Kantorovich metric between $\mathbb{P}$ and $\mathbb{Q}$ are shown in thin dotted lines in (a-c). (c) represents the behavior of $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $W(\mathbb{P}, \mathbb{Q})$ for varying d with a fixed sample size of $m = n = 250$ (see Example 5.14 for details on the choice of $\mathbb{P}$ and $\mathbb{Q}$). Error bars are obtained by replicating the experiment 20 times.

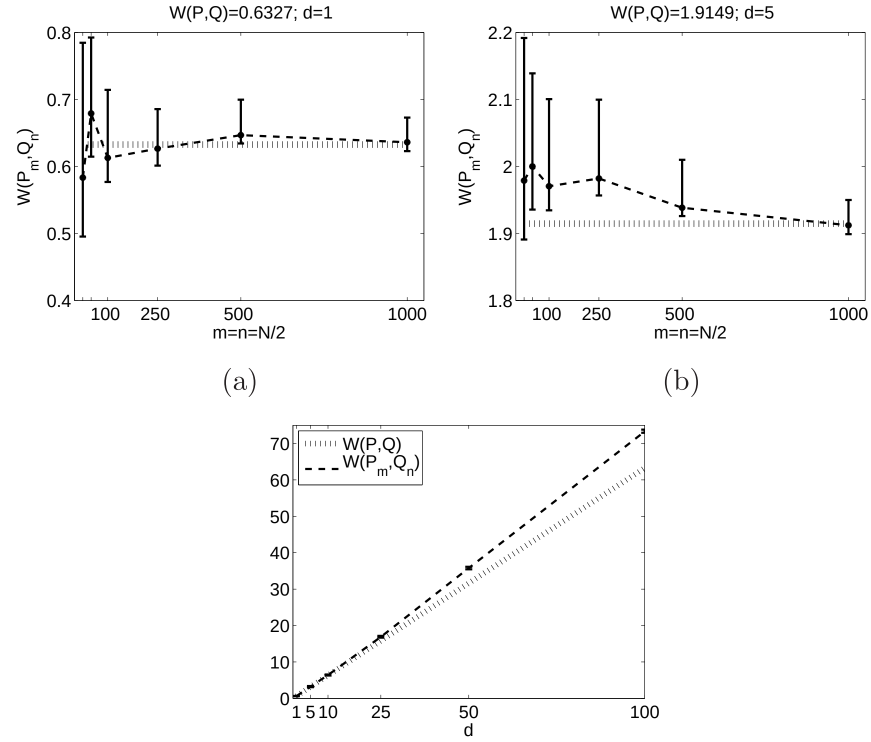
Example 5.15. *Let $\mathcal{X} = \times_{j=1}^{d}[0, c_j]$. Suppose $\mathbb{P}^{(j)}$, $\mathbb{Q}^{(j)}$ have densities*

$$p_j(x) = \frac{d\mathbb{P}^{(j)}}{dx} = \frac{\lambda_i e^{-\lambda_j x}}{1 - e^{-\lambda_j c_j}},\ q_j(x) = \frac{d\mathbb{Q}^{(j)}}{dx} = \frac{\mu_j e^{-\mu_j x}}{1 - e^{-\mu_j c_j}}$$

respectively, where $\lambda_j > 0$, $\mu_j > 0$. Note that $\mathbb{P}^{(j)}$ and $\mathbb{Q}^{(j)}$ are exponential distributions supported on $[0, c_j]$ with rate parameters λ_j and μ_j. Then, it can be shown that

$$W(\mathbb{P}^{(j)}, \mathbb{Q}^{(j)}) = \left| \frac{1}{\lambda_j} - \frac{1}{\mu_j} - \frac{c_j(e^{-\lambda_j c_j} - e^{-\mu_j c_j})}{(1 - e^{-\lambda_j c_j})(1 - e^{-\mu_j c_j})} \right|,$$

and $W(\mathbb{P}, \mathbb{Q})$ follows from (5.30).



(a)

(b)

(c)

Figure 5.2: (a-b) represent the empirical estimates of the Kantorovich metric (shown in thick dotted lines) between $\mathbb{P}$ and $\mathbb{Q}$, which are truncated exponential distributions on $\mathbb{R}^d_+$ (see Example 5.15 for details), for increasing sample size N. Here $d = 1$ in (a) and $d = 5$ in (b) with $\rho(x, y) = \|x - y\|_1$. The population values of the Kantorovich metric between $\mathbb{P}$ and $\mathbb{Q}$ are shown in thin dotted lines in (a-c). (c) represents the behavior of $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $W(\mathbb{P}, \mathbb{Q})$ for varying d with a fixed sample size of $m = n = 250$ (see Example 5.15 for details on the choice of $\mathbb{P}$ and $\mathbb{Q}$). Error bars are obtained by replicating the experiment 20 times.

Figures 5.2(a) and 5.2(b) show $W(\mathbb{P}_m, \mathbb{Q}_n)$ *(shown in thick dotted lines) for* $d = 1$ *and* $d = 5$ *respectively. Let* $\lambda = (\lambda_1, \overset{d}{\ldots}, \lambda_d)$, $\mu = (\mu_1, \overset{d}{\ldots}, \mu_d)$ *and* $c = (c_1, \overset{d}{\ldots}, c_d)$. *In Figure 5.2(a), we chose* $\lambda = (3)$, $\mu = (1)$ *and* $c = (5)$ *which gives* $W(\mathbb{P}, \mathbb{Q}) = 0.6327$. *In Figure 5.2(b), we chose* $\lambda = (3, 2, 1/2, 2, 7)$, $\mu = (1, 5, 5/2, 1, 8)$ *and* $c = (5, 6, 3, 2, 10)$, *which gives* $W(\mathbb{P}, \mathbb{Q}) = 1.9149$. *The population values* $W(\mathbb{P}, \mathbb{Q})$ *are shown in thin dotted lines in Figures 5.2(a) and 5.2(b). Figure 5.2(c) shows* $W(\mathbb{P}_m, \mathbb{Q}_n)$ *and* $W(\mathbb{P}, \mathbb{Q})$ *for various* d *with a fixed sample size of* $m = n = 250$, $\lambda = (3, 3, \overset{d}{\ldots}, 3)$, $\mu = (1, 1, \overset{d}{\ldots}, 1)$ *and* $c = (5, 5, \overset{d}{\ldots}, 5)$.

The empirical estimates in Figures 5.1 and 5.2 are obtained by drawing N

i.i.d. samples (with $m = n = N/2$) from $\mathbb{P}$ and $\mathbb{Q}$ and then solving the linear program in (5.6). It is easy to see from Figures 5.1(a,b) and 5.2(a,b) that $W(\mathbb{P}_m, \mathbb{Q}_n)$ improves with increasing sample size and that $W(\mathbb{P}_m, \mathbb{Q}_n)$ estimates $W(\mathbb{P}, \mathbb{Q})$ correctly, which therefore demonstrates the efficacy of the estimator. Figures 5.1(c) and 5.2(c) show the effect of dimensionality, d of the data on $W(\mathbb{P}_m, \mathbb{Q}_n)$. They show that at large d, the estimator has a large bias and more samples are needed to obtain better estimates. Error bars are obtained by replicating the experiment 20 times.

Empirical Estimator of $\gamma_k(\mathbb{P}, \mathbb{Q})$

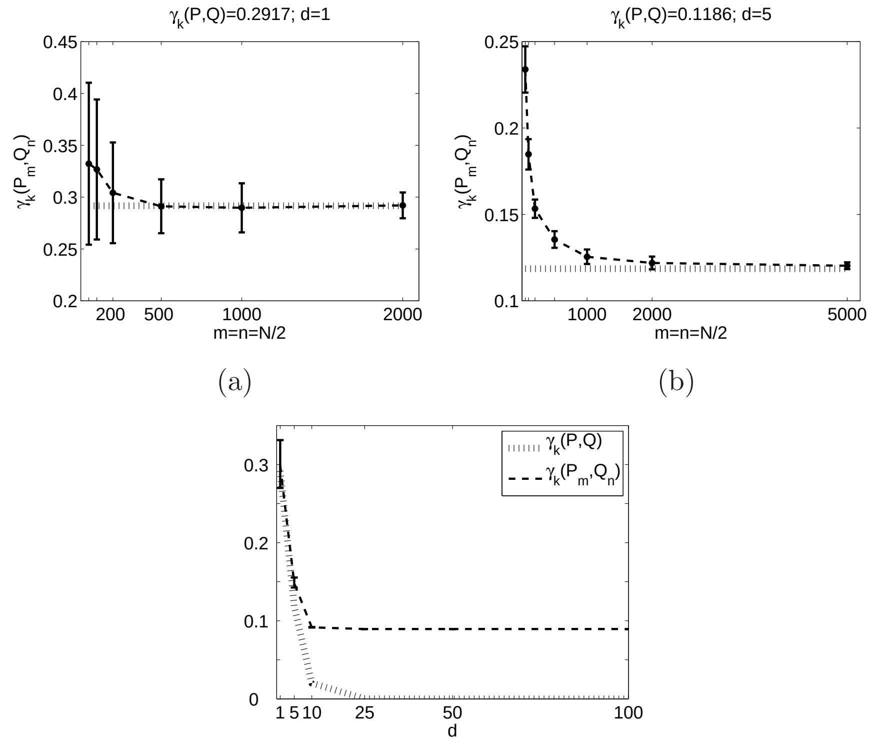
We now consider the performance of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$. Note that, although $\gamma_k(\mathbb{P}, \mathbb{Q})$ has a closed form in (3.5), exact computation is not always possible for all choices of k, $\mathbb{P}$ and $\mathbb{Q}$. In such cases, one has to resort to numerical techniques to compute the integrals in (3.5). In the following, we present two examples where we choose $\mathbb{P}$ and $\mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q})$ can be computed exactly, which is then used to verify the performance of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$. Also note that for the consistency of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$, by Proposition 5.9, we just need the kernel, k to be measurable and bounded and no assumptions on $\mathcal{X}$ are required.

Example 5.16. *Let $\mathcal{X} = \mathbb{R}^d$, $\mathbb{P} = \otimes_{j=1}^d \mathbb{P}^{(j)}$ and $\mathbb{Q} = \otimes_{j=1}^d \mathbb{Q}^{(j)}$. Suppose $\mathbb{P}^{(j)} = N(\mu_j, \sigma_j^2)$ and $\mathbb{Q}^{(i)} = N(\lambda_j, \theta_j^2)$, where $N(\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2. Let $k(x, y) = \exp(-\|x-y\|_2^2/2\tau^2)$. Clearly k is measurable and bounded. With this choice of k, $\mathbb{P}$ and $\mathbb{Q}$, γ_k in (3.5) can be computed exactly as*

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \prod_{j=1}^d \frac{\tau}{\sqrt{2\sigma_j^2 + \tau^2}} + \prod_{j=1}^d \frac{\tau}{\sqrt{2\theta_j^2 + \tau^2}} - 2\prod_{j=1}^d \frac{\tau e^{-\frac{(\mu_j - \lambda_j)^2}{2(\sigma_j^2 + \theta_j^2 + \tau^2)}}}{\sqrt{\sigma_j^2 + \theta_j^2 + \tau^2}}, \qquad (5.32)$$

as the integrals in (3.5) simply involve the convolution of Gaussian distributions.

Figures 5.3(a-b) show $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ (shown in thick dotted lines) for $d = 1$ and $d = 5$ respectively. Figure 5.3(c) shows the behavior of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ and $\gamma_k(\mathbb{P}, \mathbb{Q})$ for varying d with a fixed sample size of $m = n = 250$. Here we chose $\mu_j = 0$, $\lambda_j = 1$, $\sigma_j = \sqrt{2}$, $\theta_j = \sqrt{2}$ for all $j = 1, \ldots, d$ and $\tau = 1$. Using these values in



(c)

Figure 5.3: (a-b) represent the empirical estimates of MMD (shown in thick dotted lines) between $\mathbb{P} = N(0, 2I_d)$ and $\mathbb{Q} = N(1, 2I_d)$ with $k(x, y) = \exp(-\frac{1}{2}\|x - y\|_2^2)$, for increasing sample size N, where $d = 1$ in (a) and $d = 5$ in (b) (see Example 5.16 for details). Here $N(\mu, \sigma^2 I_d)$ represents a normal distribution with mean vector $(\mu_1, \overset{d}{\ldots}, \mu_d)$ and covariance matrix $\sigma^2 I_d$. I_d represents the $d \times d$ identity matrix. The population values of MMD are shown in thin dotted lines in (a-c). (c) represents the behavior of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ and $\gamma_k(\mathbb{P}, \mathbb{Q})$ for varying d with a fixed sample size of $m = n = 250$ (see Example 5.16 for details on the choice of $\mathbb{P}$ and $\mathbb{Q}$). Error bars are obtained by replicating the experiment 20 times.

(5.32), it is easy to check that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 5^{-d/4}(2 - 2e^{-d/10})^{1/2}$*, which is shown in thin dotted lines in Figures 5.3(a-c). We remark that an alternative estimator of* γ_k *exists which does not suffer from bias at small sample sizes: see [37].*

Example 5.17. *Let* $\mathscr{X} = \mathbb{R}_+^d$, $\mathbb{P} = \otimes_{j=1}^d \mathbb{P}^{(j)}$ *and* $\mathbb{Q} = \otimes_{j=1}^d \mathbb{Q}^{(j)}$*. Suppose* $\mathbb{P}^{(j)} = \mathrm{Exp}(1/\lambda_j)$ *and* $\mathbb{Q}^{(j)} = \mathrm{Exp}(1/\mu_j)$*, which are exponential distributions on* $\mathbb{R}_+$ *with rate parameters* $\lambda_j > 0$ *and* $\mu_j > 0$ *respectively. Suppose* $k(x, y) = \exp(-\alpha\|x - y\|_1)$*,* $\alpha > 0$*, which is a Laplacian kernel on* $\mathbb{R}^d$*. Then, it is easy to verify that*

$\gamma_k(\mathbb{P},\mathbb{Q})$ in (3.5) reduces to

$$\gamma_k^2(\mathbb{P},\mathbb{Q}) = \prod_{j=1}^d \frac{\lambda_j}{\lambda_j+\alpha} + \prod_{j=1}^d \frac{\mu_j}{\mu_j+\alpha} - 2\prod_{j=1}^d \frac{\lambda_j\mu_j(\lambda_j+\mu_j+2\alpha)}{(\lambda_j+\alpha)(\mu_j+\alpha)(\lambda_j+\mu_j)}.$$

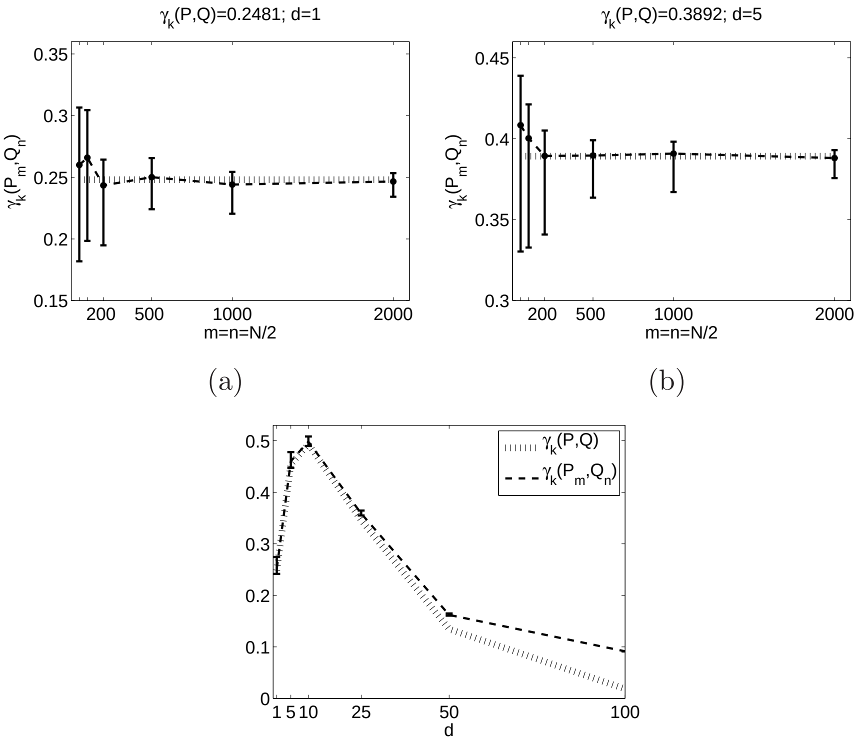
Figures 5.4(a-b) show $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ (shown in thick dotted lines) for $d=1$ and $d=5$ respectively. Figure 5.4(c) shows the dependence of $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ and $\gamma_k(\mathbb{P},\mathbb{Q})$ on d at a fixed sample size of $m=n=250$. Here, we chose $\{\lambda_j\}_{j=1}^d$ and $\{\mu_j\}_{j=1}^d$ as in Example 5.15 with $\alpha=\frac{1}{4}$, which gives $\gamma_k(\mathbb{P},\mathbb{Q})=0.2481$ for $d=1$ and 0.3892 for $d=5$, shown in thin dotted lines in Figures 5.4(a-c).

As in the case of $W(\mathbb{P}_m,\mathbb{Q}_n)$, the performance of $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ is verified by drawing N i.i.d. samples (with $m=n=N/2$) from $\mathbb{P}$ and $\mathbb{Q}$ and computing $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ in (5.13). Figures 5.3(a,b) and 5.4(a,b) show the performance of $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ for various sample sizes and some fixed d. It is easy to see that the quality of the estimate improves with increasing sample size and that $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ estimates $\gamma_k(\mathbb{P},\mathbb{Q})$ correctly. On the other hand, Figures 5.3(c) and 5.4(c) demonstrate that $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ is biased at large d and more samples are needed to obtain better estimates. As in the case of W, the error bars are obtained by replicating the experiment 20 times.

Empirical Estimator of $\beta(\mathbb{P},\mathbb{Q})$

In the case of W and γ_k, we have some closed form expression to start with (see (5.31) and (3.5)), which can be solved by numerical methods. The resulting value is then used as the baseline to test the performance of the estimators of W and γ_k. On the other hand, in the case of β, we are not aware of any such closed form expression to compute the baseline. However, it is possible to compute $\beta(\mathbb{P},\mathbb{Q})$ when $\mathbb{P}$ and $\mathbb{Q}$ are discrete distributions on $\mathcal{X}$, i.e., $\mathbb{P}=\sum_{j=1}^r \lambda_j\delta_{X_j}$, $\mathbb{Q}=\sum_{j=1}^s \mu_j\delta_{Z_j}$, where $\sum_{j=1}^r\lambda_j=1$, $\sum_{j=1}^s\mu_j=1$, $X_j,Z_j\in\mathcal{X}$, $\lambda_j\ge 0$, $\mu_j\ge 0$, $\forall\, j$. This is because, for this choice of $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\begin{aligned}\beta(\mathbb{P},\mathbb{Q}) &= \sup\left\{\sum_{j=1}^r \lambda_j f(X_j) - \sum_{j=1}^s \mu_j f(Z_i) : \|f\|_{BL}\le 1\right\}\\ &= \sup\left\{\sum_{j=1}^{r+s}\theta_j f(V_j) : \|f\|_{BL}\le 1\right\}, \qquad (5.33)\end{aligned}$$



(a)

(b)

(c)

Figure 5.4: (a-b) represent the empirical estimates of MMD (shown in thick dotted lines) between $\mathbb{P}$ and $\mathbb{Q}$, which are exponential distributions on $\mathbb{R}^d_+$ (see Example 5.17 for details), for increasing sample size N. Here $d = 1$ in (a) and $d = 5$ in (b) with $k(x, y) = \exp(-\frac{1}{4}\|x - y\|_1)$. The population values of MMD are shown in thin dotted lines in (a-c). (c) represents the behavior of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ and $\gamma_k(\mathbb{P}, \mathbb{Q})$ for varying d with a fixed sample size of $m = n = 250$ (see Example 5.17 for details on the choice of $\mathbb{P}$ and $\mathbb{Q}$). Error bars are obtained by replicating the experiment 20 times.

where $\theta = (\lambda_1, \ldots, \lambda_r, -\mu_1, \ldots, -\mu_s)$, $V = (X_1, \ldots, X_r, Z_1, \ldots, Z_s)$ with $\theta_j := (\theta)_j$ and $V_j := (V)_j$. Now, (5.33) is of the form of (5.3) and so, by Theorem 5.3, $\beta(\mathbb{P}, \mathbb{Q}) = \sum_{j=1}^{r+s} \theta_j a^\star_j$, where $\{a^\star_j\}$ solve the following linear program,

$$\begin{aligned} \max_{a_1,\ldots,a_{r+s},b,c} \quad & \sum_{j=1}^{r+s} \theta_j a_j \\ \text{s.t.} \quad & -b\,\rho(V_l, V_j) \le a_l - a_j \le b\,\rho(V_l, V_j),\ \forall\, j, l \\ & -c \le a_j \le c,\ \forall\, j \\ & b + c \le 1. \end{aligned} \tag{5.34}$$

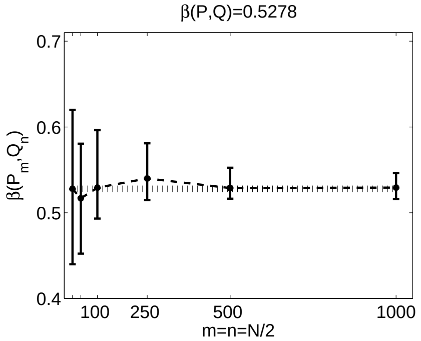


Figure 5.5: Empirical estimates of the Dudley metric (shown in a thick dotted line) between discrete distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}$ (see Example 5.18 for details), for increasing sample size N. The population value of the Dudley metric is shown in a thin dotted line. Error bars are obtained by replicating the experiment 20 times.

Therefore, for these distributions, one can compute the baseline which can then be used to verify the performance of $\beta(\mathbb{P}_m, \mathbb{Q}_n)$. In the following, we consider a simple example to demonstrate the performance of $\beta(\mathbb{P}_m, \mathbb{Q}_n)$.

Example 5.18. *Let $\mathcal{X} = \{0, 1, 2, 3, 4, 5\} \subset \mathbb{R}$, $\lambda = (\frac{1}{3}, \frac{1}{6}, \frac{1}{8}, \frac{1}{4}, \frac{1}{8})$, $\mu = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $X = (0, 1, 2, 3, 4)$ and $Z = (2, 3, 4, 5)$. With this choice, $\mathbb{P}$ and $\mathbb{Q}$ are defined as $\mathbb{P} = \sum_{j=1}^{5} \lambda_j \delta_{X_j}$ and $\mathbb{Q} = \sum_{j=1}^{4} \mu_j \delta_{Z_j}$. By solving (5.34) with $\rho(x, y) = |x - y|$, we get $\beta(\mathbb{P}, \mathbb{Q}) = 0.5278$.*

Figure 5.5 shows $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ (shown in a thick dotted line) which is computed by drawing N i.i.d. samples (with $m = n = N/2$) from $\mathbb{P}$ and $\mathbb{Q}$ and solving the linear program in (5.10). It can be seen that $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ estimates $\beta(\mathbb{P}, \mathbb{Q})$ correctly.

Since we do not know how to compute $\beta(\mathbb{P}, \mathbb{Q})$ for $\mathbb{P}$ and $\mathbb{Q}$ other than the ones we discussed here, we do not provide any other non-trivial examples to test the performance of $\beta(\mathbb{P}_m, \mathbb{Q}_n)$.

5.2.5 Empirical Estimation of Total Variation Distance

In Sections 5.2.1–5.2.4, we have derived and analyzed the empirical estimators of W, β and γ_k. Since the total variation distance,

$$TV(\mathbb{P}, \mathbb{Q}) := \sup \left\{ \int_{\mathcal{X}} f \, d(\mathbb{P} - \mathbb{Q}) \, : \, \|f\|_\infty \leq 1 \right\},$$

is also an IPM, in this section, we consider its empirical estimation and consistency analysis. Suppose $\mathcal{X}$ is a metric space. Let $TV(\mathbb{P}_m, \mathbb{Q}_n)$ be the empirical estimator of $TV(\mathbb{P}, \mathbb{Q})$. Using similar arguments as in Theorems 5.1 and 5.3, it can be shown that

$$TV(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{j=1}^{N} \widetilde{Y}_j a_j^\star,$$

where $\{a_j^\star\}_{j=1}^N$ solve the following linear program,

$$\max_{a_1,\dots,a_N} \left\{ \sum_{j=1}^{N} \widetilde{Y}_j a_j \,:\, -1 \le a_j \le 1, \, \forall\, j \right\}.$$

Now, the question is whether this estimator consistent. First note that $a_j^\star = \text{sign}(\widetilde{Y}_j)$ and therefore, $TV(\mathbb{P}_m, \mathbb{Q}_n) = 2$ for any m, n. This means for any $\mathbb{P}, \mathbb{Q}$ such that $TV(\mathbb{P}, \mathbb{Q}) < 2$, $TV(\mathbb{P}_m, \mathbb{Q}_n)$ is not a consistent estimator of $TV(\mathbb{P}, \mathbb{Q})$. Indeed $a_i^\star$, $\forall\, i$ are independent of the actual samples, $\{X_j\}_{j=1}^N$ drawn from $\mathbb{P}$ and $\mathbb{Q}$, unlike in the estimation of Kantorovich and Dudley metrics, and therefore it is not surprising that $TV(\mathbb{P}_m, \mathbb{Q}_n)$ is not a consistent estimator of $TV(\mathbb{P}, \mathbb{Q})$.

Suppose $\mathcal{X} = \mathbb{R}^d$ and let $\mathbb{P}$, $\mathbb{Q}$ be absolutely continuous w.r.t. the Lebesgue measure. Then $TV(\mathbb{P}, \mathbb{Q})$ can be consistently estimated in a strong sense using the total variation distance between the kernel density estimators of $\mathbb{P}$ and $\mathbb{Q}$. This is because if $\widetilde{\mathbb{P}}_m$ and $\widetilde{\mathbb{Q}}_n$ represent the kernel density estimators associated with $\mathbb{P}$ and $\mathbb{Q}$ respectively, then $|TV(\widetilde{\mathbb{P}}_m, \widetilde{\mathbb{Q}}_n) - TV(\mathbb{P}, \mathbb{Q})| \le TV(\widetilde{\mathbb{P}}_m, \mathbb{P}) + TV(\widetilde{\mathbb{Q}}_n, Q) \xrightarrow{a.s.} 0$ as $m, n \to \infty$ (see [21, Chapter 6] and references therein).

The issue in the empirical estimation of $TV(\mathbb{P}, \mathbb{Q})$ is that the set $\mathcal{F}_{TV} := \{f : \|f\|_\infty \le 1\}$ is too large to obtain meaningful results if no assumptions on distributions are made. On the other hand, one can choose a more manageable subset $\mathcal{F}$ of $\mathcal{F}_{TV}$ such that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) \le TV(\mathbb{P}, \mathbb{Q})$, $\forall\, \mathbb{P}, \mathbb{Q}$ and $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$ is a consistent estimator of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$. Examples of such choice of $\mathcal{F}$ include $\mathcal{F}_\beta$ and $\{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$, where the former yields the Dudley metric while the latter results in the Kolmogorov distance. The empirical estimator of the Dudley metric and its consistency have been presented in Sections 5.2.1 and 5.2.3. The empirical estimator of the Kolmogorov distance between $\mathbb{P}$ and $\mathbb{Q}$ is well studied and is strongly consistent, which simply follows from the famous Glivenko-Cantelli theorem [22, Theorem 12.4].

Since the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$ cannot be estimated consistently for all $\mathbb{P}, \mathbb{Q}$, in the following, we present two lower bounds on TV, one involving W and β and the other involving γ_k, which can be estimated consistently.

Proposition 5.19 (Lower bounds on TV)**.** *(i) Suppose $(\mathcal{X}, \rho)$ is a metric space. Then for all $\mathbb{P} \neq \mathbb{Q}$, we have*

$$TV(\mathbb{P}, \mathbb{Q}) \geq \frac{W(\mathbb{P}, \mathbb{Q})\beta(\mathbb{P}, \mathbb{Q})}{W(\mathbb{P}, \mathbb{Q}) - \beta(\mathbb{P}, \mathbb{Q})}. \tag{5.35}$$

(ii) Suppose $C := \sup_{x \in \mathcal{X}} k(x, x) < \infty$. Then

$$TV(\mathbb{P}, \mathbb{Q}) \geq \frac{\gamma_k(\mathbb{P}, \mathbb{Q})}{\sqrt{C}}. \tag{5.36}$$

Before we prove Proposition 5.19, we present an upper bound on γ_k in terms of the coupling formulation [23, Section 11.8], which is not only useful in proving Proposition 5.19(ii) (we also use it to prove Theorem 5.22) but also interesting in its own right.

Proposition 5.20 (Coupling bound)**.** *Let k be measurable and bounded on $\mathcal{X}$. Then, for any $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$,*

$$\gamma_k(\mathbb{P}, \mathbb{Q}) \leq \inf_{\mu \in \mathcal{L}(\mathbb{P}, \mathbb{Q})} \iint_{\mathcal{X}} \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}} \, d\mu(x, y), \tag{5.37}$$

where $\mathcal{L}(\mathbb{P}, \mathbb{Q})$ represents the set of all laws on $\mathcal{X} \times \mathcal{X}$ with marginals $\mathbb{P}$ and $\mathbb{Q}$.

Proof. For any $\mu \in \mathcal{L}(\mathbb{P}, \mathbb{Q})$, we have

$$\begin{aligned} \left| \int_{\mathcal{X}} f \, d(\mathbb{P} - \mathbb{Q}) \right| &= \left| \iint_{\mathcal{X}} (f(x) - f(y)) \, d\mu(x, y) \right| \\ &\leq \iint_{\mathcal{X}} |f(x) - f(y)| \, d\mu(x, y) \\ &\overset{(a)}{=} \iint_{\mathcal{X}} |\langle f, k(\cdot, x) - k(\cdot, y) \rangle_{\mathcal{H}}| \, d\mu(x, y) \\ &\overset{(b)}{\leq} \|f\|_{\mathcal{H}} \iint_{\mathcal{X}} \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}} \, d\mu(x, y), \end{aligned} \tag{5.38}$$

where we have used the reproducing property of $\mathcal{H}$ in (a) and the Cauchy-Schwartz inequality in (b). Taking the supremum over $f \in \mathcal{F}_k$ and the infimum over $\mu \in \mathcal{L}(\mathbb{P}, \mathbb{Q})$ in (5.38), where $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$, gives the result in (5.37). □

Proof of Proposition 5.19. (i) The proof is based on Lemma 5.7. Note that $\|f\|_L$, $\|f\|_{BL}$ and $\|f\|_\infty$ are convex functionals on the vector spaces $\text{Lip}(\mathcal{X},\rho)$, $BL(\mathcal{X},\rho)$ and $U(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{R} \,|\, \|f\|_\infty < \infty\}$ respectively. Similarly, $\mathbb{P}f - \mathbb{Q}f$ is a convex functional on $\text{Lip}(\mathcal{X},\rho)$, $BL(\mathcal{X},\rho)$ and $U(\mathcal{X})$, where $\mathbb{P}f := \int_{\mathcal{X}} f\,d\mathbb{P}$. Since $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}f - \mathbb{Q}f$ is not constant on $\mathcal{F}_W$, $\mathcal{F}_\beta$ and $\mathcal{F}_{TV}$. Therefore, by appropriately choosing ψ, θ, V and b in Lemma 5.7, the following sequence of inequalities are obtained.

$$\begin{aligned}
1 &= \inf\{\|f\|_{BL} : \mathbb{P}f - \mathbb{Q}f \geq \beta(\mathbb{P},\mathbb{Q}),\, f \in BL(\mathcal{X},\rho)\} \\
&\geq \inf\{\|f\|_{L} : \mathbb{P}f - \mathbb{Q}f \geq \beta(\mathbb{P},\mathbb{Q}),\, f \in BL(\mathcal{X},\rho)\} \\
&\qquad + \inf\{\|f\|_{\infty} : \mathbb{P}f - \mathbb{Q}f \geq \beta(\mathbb{P},\mathbb{Q}),\, f \in BL(\mathcal{X},\rho)\} \\
&= \frac{\beta(\mathbb{P},\mathbb{Q})}{W(\mathbb{P},\mathbb{Q})} \inf\{\|f\|_{L} : \mathbb{P}f - \mathbb{Q}f \geq W(\mathbb{P},\mathbb{Q}),\, f \in BL(\mathcal{X},\rho)\} \\
&\qquad + \frac{\beta(\mathbb{P},\mathbb{Q})}{TV(\mathbb{P},\mathbb{Q})} \inf\{\|f\|_{\infty} : \mathbb{P}f - \mathbb{Q}f \geq TV(\mathbb{P},\mathbb{Q}),\, f \in BL(\mathcal{X},\rho)\} \\
&\geq \frac{\beta(\mathbb{P},\mathbb{Q})}{W(\mathbb{P},\mathbb{Q})} \inf\{\|f\|_{L} : \mathbb{P}f - \mathbb{Q}f \geq W(\mathbb{P},\mathbb{Q}),\, f \in \text{Lip}(\mathcal{X},\rho)\} \\
&\qquad + \frac{\beta(\mathbb{P},\mathbb{Q})}{TV(\mathbb{P},\mathbb{Q})} \inf\{\|f\|_{\infty} : \mathbb{P}f - \mathbb{Q}f \geq TV(\mathbb{P},\mathbb{Q}),\, f \in U(\mathcal{X})\} \\
&= \frac{\beta(\mathbb{P},\mathbb{Q})}{W(\mathbb{P},\mathbb{Q})} + \frac{\beta(\mathbb{P},\mathbb{Q})}{TV(\mathbb{P},\mathbb{Q})},
\end{aligned}$$

which gives (5.35).

(ii) To prove (5.36), we use the coupling formulation for TV [49, p. 19] given by

$$TV(\mathbb{P},\mathbb{Q}) = 2 \inf_{\mu \in \mathcal{L}(\mathbb{P},\mathbb{Q})} \mu(X \neq Y), \tag{5.39}$$

where $\mathcal{L}(\mathbb{P},\mathbb{Q})$ is the set of all measures on $\mathcal{X} \times \mathcal{X}$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. Here, X and Y are distributed as $\mathbb{P}$ and $\mathbb{Q}$ respectively. Let $\lambda \in \mathcal{L}(\mathbb{P},\mathbb{Q})$ and $f \in \mathcal{H}$. Consider

$$\begin{aligned}
\|k(\cdot,x) - k(\cdot,y)\|_{\mathcal{H}} &\leq \mathbb{1}_{x\neq y}\|k(\cdot,x) - k(\cdot,y)\|_{\mathcal{H}} \\
&\leq \mathbb{1}_{x\neq y}\left[\|k(\cdot,x)\|_{\mathcal{H}} + \|k(\cdot,y)\|_{\mathcal{H}}\right] \\
&= \mathbb{1}_{x\neq y}\left[\sqrt{k(x,x)} + \sqrt{k(y,y)}\right] \\
&\leq 2\sqrt{C}\mathbb{1}_{x\neq y}.
\end{aligned} \tag{5.40}$$

Using (5.40) in (5.37) yields (5.36) through (5.39). □

Remark 5.21. (i) *As mentioned before, a simple lower bound on TV can be obtained as $TV(\mathbb{P},\mathbb{Q}) \ge \beta(\mathbb{P},\mathbb{Q})$, $\forall\,\mathbb{P},\mathbb{Q}\in M^1_+(\mathcal{X})$. It is easy to see that the bound in (5.35) is tighter as $\frac{W(\mathbb{P},\mathbb{Q})\beta(\mathbb{P},\mathbb{Q})}{W(\mathbb{P},\mathbb{Q})-\beta(\mathbb{P},\mathbb{Q})} \ge \beta(\mathbb{P},\mathbb{Q})$ with equality if and only if $\mathbb{P}=\mathbb{Q}$.*

(ii) *The bounds in (5.35) and (5.36) translate as lower bounds on the KL-divergence through Pinsker's inequality: $TV^2(\mathbb{P},\mathbb{Q}) \le 2\,D_{t\log t}(\mathbb{P},\mathbb{Q})$, $\forall\,\mathbb{P},\mathbb{Q}\in M^1_+(\mathcal{X})$. See [25] and references therein for more refined bounds between TV and KL-divergence. Therefore, using these bounds, one can obtain a consistent estimate of a lower bound on TV and KL-divergence. The bounds in (5.35) and (5.36) also translate to lower bounds on other distance measures on probabilities. See [33] for a detailed discussion on the relation between various metrics.*

5.3 Metrization of the Weak Topology

As motivated in Section 5.1, an important question to consider that is useful both in theory and practice would be: "How strong or weak is γ_k related to other metrics on $M^1_+(\mathcal{X})$?" This question is addressed in Theorem 5.22, where we compare γ_k to other metrics on $M^1_+(\mathcal{X})$ like the Dudley metric (β), Wasserstein distance (W), total variation distance (TV), and show that γ_k is weaker than all these metrics (see footnote 15 for the definition of "strong" and "weak" metrics). Since γ_k is weaker than β, which is known to induce a topology on $M^1_+(\mathcal{X})$ that coincides with the standard topology on $M^1_+(\mathcal{X})$, called the weak* (weak-star) topology (usually called the weak topology in probability theory and from now on we use these terms interchangeably), this naturally leads to the question, "For what k does the topology induced by γ_k coincide with the weak topology?" Although we arrived at this question motivated by an application, this question on its own is theoretically interesting and important in probability theory, especially in proving central limit theorems. In Theorem 5.24 we show that c_0-universal kernels metrize the weak*-topology on $M^1_+(\mathcal{X})$, assuming $\mathcal{X}$ to be a Polish and LCH space.

First, we start with some preliminaries. The *weak topology* on $M^1_+(\mathcal{X})$ is the weakest topology such that the map $\mathbb{P}\mapsto\int_{\mathcal{X}} f\,d\mathbb{P}$ is continuous for all $f\in C_b(\mathcal{X})$. For a metric space $(\mathcal{X},\rho)$, a sequence $\mathbb{P}_n$ of probability measures is said to *converge weakly* to $\mathbb{P}$, written as $\mathbb{P}_n \overset{w}{\to} \mathbb{P}$, if and only if $\int_{\mathcal{X}} f\, d\mathbb{P}_n \to \int_{\mathcal{X}} f\, d\mathbb{P}$ for every $f \in C_b(\mathcal{X})$. A metric γ on $M_+^1(\mathcal{X})$ is said to *metrize* the weak topology if the topology induced by γ coincides with the weak topology, which is defined as follows: if, for $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots \in M_+^1(\mathcal{X})$, $(\mathbb{P}_n \overset{w}{\to} \mathbb{P} \Leftrightarrow \gamma(\mathbb{P}_n, \mathbb{P}) \overset{n\to\infty}{\longrightarrow} 0)$ holds, then the topology induced by γ coincides with the weak topology.

In the following, we collect well-known results on the relation between various metrics on $M_+^1(\mathcal{X})$, which will be helpful in understanding the behavior of these metrics, both with respect to each other and to ours. Let $(\mathcal{X}, \rho)$ be a separable metric space. The *Prohorov metric* on $(\mathcal{X}, \rho)$, defined as

$$\varsigma(\mathbb{P}, \mathbb{Q}) := \inf\{\epsilon > 0 : \mathbb{P}(A) \leq \mathbb{Q}(A^\epsilon) + \epsilon, \forall\, \text{Borel sets } A\},$$

metrizes the weak topology on $M_+^1(\mathcal{X})$ [23, Theorem 11.3.3], where $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X})$ and $A^\epsilon := \{y \in \mathcal{X} : \rho(x, y) < \epsilon \text{ for some } x \in A\}$. Since the Dudley metric is related to the Prohorov metric as

$$\frac{1}{2}\beta(\mathbb{P}, \mathbb{Q}) \leq \varsigma(\mathbb{P}, \mathbb{Q}) \leq 2\sqrt{\beta(\mathbb{P}, \mathbb{Q})},$$

it also metrizes the weak topology on $M_+^1(\mathcal{X})$ [23, Theorem 11.3.3]. W and TV are related to the Prohorov metric as [33, Theorem 2]

$$\varsigma^2(\mathbb{P}, \mathbb{Q}) \leq W(\mathbb{P}, \mathbb{Q}) \leq (\text{diam}(\mathcal{X}) + 1)\varsigma(\mathbb{P}, \mathbb{Q}), \tag{5.41}$$

and

$$\varsigma(\mathbb{P}, \mathbb{Q}) \leq \frac{1}{2} TV(\mathbb{P}, \mathbb{Q}).$$

This means W and TV are stronger than ς, while W and ς are equivalent (i.e., induce the same topology) when $\mathcal{X}$ is bounded. By Theorem 4 in [33], TV and W are related as

$$W(\mathbb{P}, \mathbb{Q}) \leq \frac{\text{diam}(\mathcal{X})}{2} TV(\mathbb{P}, \mathbb{Q}),$$

which means W and TV are comparable if $\mathcal{X}$ is bounded. See [71, Chapter 19, Theorem 2.4] and [33] for further detail on the relationship between various metrics on $M_+^1(\mathcal{X})$.

Let us now consider a sequence of probability measures on $\mathbb{R}$, $\mathbb{P}_n := \frac{1}{n}\delta_n + \left(1 - \frac{1}{n}\right)\delta_0$ and let $\mathbb{P} := \delta_0$. It can be shown that $\beta(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$ which means $\mathbb{P}_n \overset{w}{\to} \mathbb{P}$, while $W(\mathbb{P}_n,\mathbb{P})=1$ and $TV(\mathbb{P}_n,\mathbb{P})=1$ for all n. $\gamma_k(\mathbb{P}_n,\mathbb{P})$ can be computed as

$$\gamma_k^2(\mathbb{P}_n,\mathbb{P})=\frac{1}{n^2}\iint_{\mathbb{R}} k(x,y)\,d(\delta_0-\delta_n)(x)\,d(\delta_0-\delta_n)(y)=\frac{k(0,0)+k(n,n)-2k(0,n)}{n^2}.$$

If k is, e.g., a Gaussian, Laplacian or inverse multiquadratic, then $\gamma_k(\mathbb{P}_n,\mathbb{P})\to 0$ as $n\to\infty$. This example shows that γ_k is weaker than W and TV. It also shows that, for certain choices of k, γ_k behaves similarly to β, which leads to the aforementioned questions: What is the general behavior of γ_k compared to other metrics? When does γ_k metrize the weak topology on $M^1_+(\mathcal{X})$? In other words, depending on k, how weak or strong is γ_k compared to other metrics on $M^1_+(\mathcal{X})$? Understanding the answer to these questions is important both in theory and practice. If k is such that γ_k metrizes the weak topology on $M^1_+(\mathcal{X})$, then it can be used as a theoretical tool in probability theory, similar to the Prohorov and Dudley metrics. On the other hand, the answer to these questions is critical in applications as it will have a bearing on the choice of kernels to be used.

With the above motivation, we first compare γ_k to β, W and TV. Since β is equivalent to ς and TV is related to KL-divergence (through Pinsker's inequality), we do not compare γ_k to ς and KL-divergence.

Theorem 5.22 (Comparison of γ_k to β, W and TV)**.** *Assume* $\sup_{x\in\mathcal{X}} k(x,x)\le C<\infty$, *where k is measurable on $\mathcal{X}$. Let*

$$\rho(x,y)=\|k(\cdot,x)-k(\cdot,y)\|_{\mathcal{H}}. \tag{5.42}$$

Then, for any $\mathbb{P},\mathbb{Q}\in M^1_+(\mathcal{X})$,

(i) $\gamma_k(\mathbb{P},\mathbb{Q})\le W(\mathbb{P},\mathbb{Q})\le\sqrt{\gamma_k^2(\mathbb{P},\mathbb{Q})+4C}$ *if $(\mathcal{X},\rho)$ is separable.*

(ii) $\frac{\gamma_k(\mathbb{P},\mathbb{Q})}{(1+\sqrt{C})}\le\beta(\mathbb{P},\mathbb{Q})\le 2(\gamma_k^2(\mathbb{P},\mathbb{Q})+4C)^{\frac{1}{3}}$ *if $(\mathcal{X},\rho)$ is separable.*

(iii) $\gamma_k(\mathbb{P},\mathbb{Q})\le\sqrt{C}\,TV(\mathbb{P},\mathbb{Q})$.

Proof. (i) When $(\mathcal{X},\rho)$ is separable, $W(\mathbb{P},\mathbb{Q})=W_1(\mathbb{P},\mathbb{Q})$ has a coupling formulation [23, p. 420], given as

$$W(\mathbb{P},\mathbb{Q})=\inf_{\mu\in\mathcal{L}(\mathbb{P},\mathbb{Q})}\iint_{\mathcal{X}}\rho(x,y)\,d\mu(x,y), \tag{5.43}$$

where $\mathbb{P},\mathbb{Q}\in\{\mathbb{P}\in M_+^1(\mathcal{X}):\int_{\mathcal{X}}\rho(x,y)\,d\mathbb{P}(y)<\infty,\,\forall\,x\in\mathcal{X}\}$. In our case $\rho(x,y)=\Vert k(\cdot,x)-k(\cdot,y)\Vert_{\mathcal{H}}$. In addition, $(\mathcal{X},\rho)$ is bounded, which means (5.43) holds for all $\mathbb{P},\mathbb{Q}\in M_+^1(\mathcal{X})$. The lower bound therefore follows from (5.37). The upper bound can be obtained as follows. Consider $W(\mathbb{P},\mathbb{Q})=\inf_{\mu\in\mathcal{L}(\mathbb{P},\mathbb{Q})}\iint_{\mathcal{X}}\Vert k(\cdot,x)-k(\cdot,y)\Vert_{\mathcal{H}}\,d\mu(x,y)$, which can be bounded as

$$\begin{aligned}
W(\mathbb{P},\mathbb{Q}) &\le \iint_{\mathcal{X}}\Vert k(\cdot,x)-k(\cdot,y)\Vert_{\mathcal{H}}\,d\mathbb{P}(x)\,d\mathbb{Q}(y)\\
&\overset{(a)}{\le}\left[\iint_{\mathcal{X}}\Vert k(\cdot,x)-k(\cdot,y)\Vert^2_{\mathcal{H}}\,d\mathbb{P}(x)\,d\mathbb{Q}(y)\right]^{\frac{1}{2}}\\
&\le\left[\int_{\mathcal{X}}k(x,x)\,d(\mathbb{P}+\mathbb{Q})(x)-2\iint_{\mathcal{X}}k(x,y)\,d\mathbb{P}(x)\,d\mathbb{Q}(y)\right]^{\frac{1}{2}}\\
&\le\left[\gamma_k^2(\mathbb{P},\mathbb{Q})+\iint_{\mathcal{X}}(k(x,x)-k(x,y))\,d(\mathbb{P}\otimes\mathbb{P}+\mathbb{Q}\otimes\mathbb{Q})(x,y)\right]^{\frac{1}{2}}\\
&\le\sqrt{\gamma_k^2(\mathbb{P},\mathbb{Q})+4C}, \qquad (5.44)
\end{aligned}$$

where we have used Jensen's inequality [26, p. 109] in (a).

(ii) Let $\mathcal{F}:=\{f:\Vert f\Vert_{\mathcal{H}}<\infty\}$ and $\mathcal{G}:=\{f:\Vert f\Vert_{BL}<\infty\}$. For $f\in\mathcal{F}$, we have

$$\begin{aligned}
\Vert f\Vert_{BL}&=\sup_{x\ne y}\frac{|f(x)-f(y)|}{\rho(x,y)}+\sup_{x\in\mathcal{X}}|f(x)|\\
&=\sup_{x\ne y}\frac{|\langle f,k(\cdot,x)-k(\cdot,y)\rangle_{\mathcal{H}}|}{\Vert k(\cdot,x)-k(\cdot,y)\Vert_{\mathcal{H}}}+\sup_{x\in\mathcal{X}}|\langle f,k(\cdot,x)\rangle_{\mathcal{H}}|\\
&\le(1+\sqrt{C})\Vert f\Vert_{\mathcal{H}}<\infty,
\end{aligned}$$

which implies $f\in\mathcal{G}$ and, therefore, $\mathcal{F}\subset\mathcal{G}$. Define $\mathbb{P}f:=\int_{\mathcal{X}}f\,d\mathbb{P}$. For any $\mathbb{P},\mathbb{Q}\in M_+^1(\mathcal{X})$,

$$\begin{aligned}
\gamma_k(\mathbb{P},\mathbb{Q})&=\sup\{|\mathbb{P}f-\mathbb{Q}f|:f\in\mathcal{F}_k\}\\
&\le\sup\{|\mathbb{P}f-\mathbb{Q}f|:\Vert f\Vert_{BL}\le(1+\sqrt{C}),\ f\in\mathcal{F}\}\\
&\le\sup\{|\mathbb{P}f-\mathbb{Q}f|:\Vert f\Vert_{BL}\le(1+\sqrt{C}),\ f\in\mathcal{G}\}\\
&=(1+\sqrt{C})\beta(\mathbb{P},\mathbb{Q}).
\end{aligned}$$

The upper bound is obtained as follows. For any $\mathbb{P},\mathbb{Q}\in M_+^1(\mathcal{X})$, by Markov's inequality [26, Theorem 6.17], for all $\epsilon>0$, we have

$$\epsilon^2\mu(\Vert k(\cdot,X)-k(\cdot,Y)\Vert_{\mathcal{H}}>\epsilon)\le\iint_{\mathcal{X}}\Vert k(\cdot,x)-k(\cdot,y)\Vert^2_{\mathcal{H}}\,d\mu(x,y),$$

where X and Y are distributed as $\mathbb{P}$ and $\mathbb{Q}$ respectively. Choose ϵ such that $\epsilon^3 = \iint_{\mathcal{X}} \|k(\cdot,x) - k(\cdot,y)\|^2_{\mathcal{H}}\,d\mu(x,y)$, such that $\mu(\|k(\cdot,X) - k(\cdot,Y)\|_{\mathcal{H}} > \epsilon) \le \epsilon$. From the proof of Theorem 11.3.5 in [23], when $(\mathcal{X},\rho)$ is separable, we have

$$\mu(\rho(X,Y) \ge \epsilon) < \epsilon \Rightarrow \varsigma(\mathbb{P},\mathbb{Q}) \le \epsilon,$$

which implies that

$$\begin{aligned}\varsigma(\mathbb{P},\mathbb{Q}) &\le \left(\inf_{\mu\in\mathcal{L}(\mathbb{P},\mathbb{Q})} \iint_{\mathcal{X}} \|k(\cdot,x) - k(\cdot,y)\|^2_{\mathcal{H}}\,d\mu(x,y)\right)^{\frac{1}{3}} \\ &\le \left(\iint_{\mathcal{X}} \|k(\cdot,x) - k(\cdot,y)\|^2_{\mathcal{H}}\,d\mathbb{P}(x)\,d\mathbb{Q}(y)\right)^{\frac{1}{3}} \overset{(b)}{\le} \left(\gamma^2_k(\mathbb{P},\mathbb{Q}) + 4C\right)^{\frac{1}{3}},\end{aligned}$$

where (b) follows from (5.44). The result follows from (5.41).

(iii) See the proof of Proposition 5.19(ii). □

Remark 5.23. *(a) First, note that, since k is bounded, $(\mathcal{X},\rho)$ is a bounded metric space. In addition, the metric, ρ, which depends on the kernel as in (5.42), is a Hilbertian metric (see footnote 7) on $\mathcal{X}$. A popular example of such a metric is $\rho(x,y) = \|x-y\|_2$, which can be obtained by choosing $\mathcal{X}$ to be a compact subset of $\mathbb{R}^d$ and $k(x,y) = \langle x,y\rangle$.*

(b) Theorem 5.22 shows that γ_k is weaker than β, W and TV for the assumptions being made on k and ρ. Note that the result holds irrespective of whether or not the kernel is characteristic, as we have not assumed anything about the kernel except it being measurable and bounded. Also, it is important to remember that the result holds when ρ is Hilbertian, as mentioned in (5.42) (see Remark 5.23(d)).

(c) Apart from showing that γ_k is weaker than β, W and TV, the result in Theorem 5.22 can be used to bound these metrics in terms of γ_k. For β, which is primarily of theoretical interest, we do not know a closed form expression, and likewise a closed form expression for W is known only for $\mathbb{R}$ [84]—see footnote 19. Since γ_k is easy to compute (see (3.5) and (3.6)), bounds on W can be obtained from Theorem 5.22 in terms of γ_k. A closed form expression for TV is available if $\mathbb{P}$ and $\mathbb{Q}$ have Radon-Nikodym derivatives w.r.t. a σ-finite measure. However, from Theorem 5.22, a simple lower bound can be obtained on TV in terms of γ_k

for any $\mathbb{P}, \mathbb{Q} \in M^1_+(\mathcal{X})$.

(d) In Theorem 5.22, the kernel is fixed and ρ *is defined as in (5.42), which is a Hilbertian metric. On the other hand, suppose a Hilbertian metric* ρ *is given. Then the associated kernel* k *can be obtained from* ρ *[8, Chapter 3, Lemma 2.1] as*

$$k(x,y) = \frac{1}{2}[\rho^2(x,x_0) + \rho^2(y,x_0) - \rho^2(x,y)],\ x,y,x_0 \in \mathcal{X},$$

which can then be used to compute γ_k.

The discussion so far has been devoted to relating γ_k to β, W and TV to understand the strength or weakness of γ_k w.r.t. these metrics. In a next step, we address the second question of when γ_k metrizes the weak topology on $M^1_+(\mathcal{X})$. This question would have been answered had the result in Theorem 5.22 shown that under some conditions on k, γ_k is equivalent to β. Since Theorem 5.22 does not help in this regard, we address the question as follows.

Theorem 5.24. *Let* $\mathcal{X}$ *be an LCH space (that is also Polish) and* k *be* c_0*-universal. Then, the topology induced by* γ_k *coincides with the weak topology on* $M^1_+(\mathcal{X})$.

Proof. We need to show that for measures $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots \in M^1_+(\mathcal{X})$, $\mathbb{P}_n \overset{w}{\to} \mathbb{P}$ if and only if $\gamma_k(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$. Define $\mathbb{P}f := \int_{\mathcal{X}} f\, d\mathbb{P}$. One direction is trivial as $\mathbb{P}_n \overset{w}{\to} \mathbb{P}$, i.e., $\mathbb{P}_n f \to \mathbb{P}f$, $\forall\, f \in C_b(\mathcal{X})$ implies $\mathbb{P}_n f \to \mathbb{P}f$, $\forall\, f \in \mathcal{H}$ and therefore $\gamma_k(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$. We prove the other direction as follows. Since k is c_0-universal, $\mathcal{H}$ is dense in $C_0(\mathcal{X})$ w.r.t. $\|\cdot\|_\infty$, i.e., for any $f \in C_0(\mathcal{X})$ and every $\epsilon > 0$, there exists a $g \in \mathcal{H}$ such that $\|f - g\|_\infty \le \epsilon$. Therefore,

$$\begin{aligned} |\mathbb{P}_n f - \mathbb{P}f| &= |\mathbb{P}_n(f-g) + \mathbb{P}(g-f) + (\mathbb{P}_n g - \mathbb{P}g)| \\ &\le \mathbb{P}_n|f-g| + \mathbb{P}|f-g| + |\mathbb{P}_n g - \mathbb{P}g| \\ &\le 2\epsilon + |\mathbb{P}_n g - \mathbb{P}g| \\ &\overset{(a)}{\le} 2\epsilon + \|g\|_{\mathcal{H}}\gamma_k(\mathbb{P}_n, \mathbb{P}), \end{aligned}$$

where we used Proposition 3.2 in (a) and $\mathbb{P}f := \int_{\mathcal{X}} f\, d\mathbb{P}$. Since $\gamma_k(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$ and ϵ is arbitrary, $|\mathbb{P}_n f - \mathbb{P}f| \to 0$ for any $f \in C_0(\mathcal{X})$. The result follows from [8, Corollary 4.3], which says that if $\mathbb{P}_n f \to \mathbb{P}f$, $\forall\, f \in C_0(\mathcal{X})$, then $\mathbb{P}_n f \to \mathbb{P}f$, $\forall\, f \in C_b(\mathcal{X})$, i.e., $\mathbb{P}_n \overset{w}{\to} \mathbb{P}$. □

Theorem 5.24 shows that if k is c_0-universal, then MMD induces the same topology as induced by the Prohorov and Dudley metrics and therefore is equivalent to both these metrics. This means that, although k being characteristic is sufficient to guarantee γ_k being a metric, a stronger condition on k, i.e., k being c_0-universal is required for γ_k to metrize the weak topology on $M^1_+(\mathcal{X})$.

5.4 Discussion

In this chapter, we have investigated the benefits and drawbacks of MMD in comparison to ϕ-divergences and other IPMs. We showed that: (a) the empirical estimator of MMD is easy to implement as it can be obtained in a closed form compared to those of KL-divergence, Kantorovich and Dudley metrics and (b) the empirical estimator of MMD has a better rate of convergence compared to those of these other metrics, though all these estimators are strongly consistent. On the other hand, MMD is weaker than the Kantorovich metric and KL-divergence which can be advantageous or disadvantageous depending on the problem at hand.

There are couple of interesting problems yet to be explored in connection with this work: (i) While we used empirical estimators for the estimation of W, β and γ_k, it is not clear whether the obtained rates are optimal in the minimax sense as the minimax rate for estimating W, β and γ_k has not been established and (ii) estimation of the Fortet-Mourier metric along with the convergence analysis.

Bibliographic Notes

This chapter is based on joint work with Kenji Fukumizu, Arthur Gretton, Gert Lanckriet and Bernhard Schölkopf, which appeared in [75, 76, 78]. The dissertation author was the primary investigator and author of these papers.

6 Choice of Characteristic Kernel and Two-Sample Test

In Chapter 1 and Section 5.2.2, we have impressed on the importance of using characteristic kernels in applications like hypothesis testing, binary classification, etc., the various characterizations for which are provided in Chapters 3 and 4. Let us consider the Gaussian kernel on $\mathbb{R}^d$, i.e., $k_\sigma(x,y) = \exp(-\sigma\|x-y\|_2^2)$, $\sigma \in \mathbb{R}_+$, which is shown to be characteristic for any $\sigma \in \mathbb{R}_+$, the bandwidth parameter. This means $\{k_\sigma : \sigma \in \mathbb{R}_+\}$ is the family of Gaussian kernels and $\{\gamma_{k_\sigma} : \sigma \in \mathbb{R}_+\}$ is the family of metrics on $M_+^1(\mathbb{R}^d)$ indexed by the kernel parameter, σ. However, in practice, one would prefer a single number that defines the distance between $\mathbb{P}$ and $\mathbb{Q}$. The question therefore to be addressed is how to choose appropriate σ. The choice of σ has important implications on the statistical aspect of γ_{k_σ}. Note that as $\sigma \to 0$, $k_\sigma \to 1$ and as $\sigma \to \infty$, $k_\sigma \to 0$ a.e., which means $\gamma_{k_\sigma}(\mathbb{P},\mathbb{Q}) \to 0$ as $\sigma \to 0$ or $\sigma \to \infty$ for all $\mathbb{P},\mathbb{Q} \in M_+^1(\mathbb{R}^d)$ (this behavior is also exhibited by $k_\sigma(x,y) = \exp(-\sigma\|x-y\|_1)$ and $k_\sigma(x,y) = \sigma^2/(\sigma^2 + \|x-y\|_2^2)$, which are also characteristic). This means choosing *sufficiently small* or *sufficiently large* σ (depending on $\mathbb{P}$ and $\mathbb{Q}$) makes $\gamma_{k_\sigma}(\mathbb{P},\mathbb{Q})$ arbitrarily small. Therefore, σ has to be chosen appropriately in applications to effectively distinguish between $\mathbb{P}$ and $\mathbb{Q}$.

This chapter is organized as follows. In Section 6.1, we propose a generalization of γ_k (called as the generalized MMD), yielding a new distance measure between $\mathbb{P}$ and $\mathbb{Q}$, which addresses the questions raised above. Since the metric has to be applicable in practice, we show in Section 6.2 that an empirical estimator of the generalized MMD based on finite samples is strongly consistent and establish its rate of convergence. Finally, in Section 6.3, we provide a simple experimental demonstration that the generalized MMD can be applied in practice to the problem of homogeneity testing. Specifically, we show that when two distributions differ on particular length scales, the kernel selected by the generalized MMD is appropriate to this difference, and the resulting hypothesis test outperforms the heuristic kernel choice employed in earlier studies [37].

6.1 Generalizing the MMD for Classes of Characteristic Kernels

Let us consider the following modification to γ_k, which yields a pseudometric on $M^1_+(\mathcal{X})$,

$$\gamma(\mathbb{P},\mathbb{Q}) = \sup_{k\in\mathcal{K}} \gamma_k(\mathbb{P},\mathbb{Q}) = \sup_{k\in\mathcal{K}} \left\Vert \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,x)\,d\mathbb{Q}(x)\right\Vert_{\mathcal{H}}, \tag{6.1}$$

where $\mathcal{K}$ is a family of pd kernels. Note that $\gamma(\mathbb{P},\mathbb{Q})$ is the maximal RKHS distance between $\mathbb{P}$ and $\mathbb{Q}$ over the family, $\mathcal{K}$, where examples of $\mathcal{K}$ include:

(a_1) $\mathcal{K}_g := \left\{e^{-\sigma\Vert x-y\Vert_2^2},\, x,y\in\mathbb{R}^d \,:\, \sigma\in\mathbb{R}_+\right\}$.

(a_2) $\mathcal{K}_l := \left\{e^{-\sigma\Vert x-y\Vert_1},\, x,y\in\mathbb{R}^d \,:\, \sigma\in\mathbb{R}_+\right\}$.

(a_3) $\mathcal{K}_\psi := \left\{e^{-\sigma\psi(x,y)},\, x,y\in\mathcal{X} \,:\, \sigma\in\mathbb{R}_+\right\}$, where $-\psi$ is a conditionally pd kernel on $\mathcal{X}\times\mathcal{X}$.

(a_4) $\mathcal{K}_{rbf} := \left\{\int_0^\infty e^{-\lambda\Vert x-y\Vert_2^2}\,d\mu_\sigma(\lambda), x,y\in\mathbb{R}^d,\, \mu_\sigma\in\widetilde{M_b^+}(\mathbb{R}_+) \,:\, \sigma\in\mathbb{R}_+\right\}$, where $\widetilde{M_b^+}(\mathbb{R}_+) := \left\{\mu\in M_b^+(\mathbb{R}_+)\,|\,\text{supp}(\mu)\neq\{0\}\right\}$.

(a_5) $\mathcal{K}_{lin} := \left\{k_\lambda = \sum_{j=1}^s \lambda_j k_j \,|\, k_\lambda \text{ is pd}, \sum_{j=1}^s \lambda_j = 1\right\}$, which is the linear combination of pd kernels $\{k_j\}_{j=1}^s$.

(a_6) $\mathcal{K}_{con} := \left\{k_\lambda = \sum_{j=1}^s \lambda_j k_j \,|\, \lambda_j\geq 0, \forall\, j, \sum_{j=1}^s \lambda_j = 1\right\}$, which is the convex combination of pd kernels $\{k_j\}_{j=1}^s$.

It is easy to check that if any $k\in\mathcal{K}$ is characteristic, then γ is a metric on $M^1_+(\mathcal{X})$. From the definition of $\gamma(\mathbb{P},\mathbb{Q})$, it is clear that we use $k^* = \arg\sup\{\gamma_k(\mathbb{P},\mathbb{Q}) : k\in$

$\mathcal{K}\}$ to compute $\gamma_{k^*}(\mathbb{P},\mathbb{Q})$, which means for $\mathcal{K}$ in (a_1)–(a_4), we choose

$$\sigma^* = \arg\sup\{\gamma_{k_\sigma}(\mathbb{P},\mathbb{Q}) \,:\, \sigma \in \mathbb{R}_+\}$$

to compute $\gamma_{k_{\sigma^*}}$. Therefore, the definition of $\gamma(\mathbb{P},\mathbb{Q})$ in (6.1) addresses the questions that we raised before.

The proposal of $\gamma(\mathbb{P},\mathbb{Q})$ in (6.1) can be motivated by the connection that we have established in Section 5.2.2 between γ_k and the Parzen window classifier. Since the Parzen window classifier depends on the kernel, k, one can propose to learn the kernel like in support vector machines [46], wherein the kernel is chosen such that $R^L_{\mathcal{F}_k}$ in Theorem 5.5 is minimized over $k \in \mathcal{K}$, i.e., $\inf_{k\in\mathcal{K}} R^L_{\mathcal{F}_k} = -\sup_{k\in\mathcal{K}}\gamma_k(\mathbb{P},\mathbb{Q}) = -\gamma(\mathbb{P},\mathbb{Q})$. A similar motivation for γ can be provided based on Proposition 5.6 as learning the kernel in a hard-margin support vector machine by maximizing its margin.

The idea and validity behind the proposal of γ in (6.1) can also be understood from a Bayesian perspective, where we define a nonnegative finite measure η over $\mathcal{K}$, and average γ_k over that measure, i.e.,

$$\alpha(\mathbb{P},\mathbb{Q}) := \int_{\mathcal{K}} \gamma_k(\mathbb{P},\mathbb{Q})\, d\eta(k).$$

This also yields a pseudometric on $M^1_+(\mathscr{X})$. That said, $\alpha(\mathbb{P},\mathbb{Q}) \le \eta(\mathcal{K})\gamma(\mathbb{P},\mathbb{Q})$ for any $\mathbb{P}$, $\mathbb{Q}$, which means that, if $\mathbb{P}$ and $\mathbb{Q}$ can be distinguished by α, then they can be distinguished by γ, but not vice-versa. In this sense, γ is stronger than α and therefore studying γ makes sense. One further complication with the Bayesian approach is in defining a sensible η over $\mathcal{K}$. Note that γ_{k_0} can be obtained by defining $\eta(k) = \delta(k - k_0)$ in $\alpha(\mathbb{P},\mathbb{Q})$.

At this point, we briefly discuss the issue of normalized vs. unnormalized kernel families, $\mathcal{K}$ in (6.1). We say a translation invariant kernel, k on $\mathbb{R}^d$ is normalized if $\int_{\mathbb{R}^d}\psi(y)\,dy = c$ (some positive constant independent of the kernel parameter), where $k(x,y) = \psi(x-y)$. $\mathcal{K}$ is a normalized kernel family if every kernel in $\mathcal{K}$ is normalized. If $\mathcal{K}$ is not normalized, we say it is unnormalized. For example, it is easy to see that $\mathcal{K}_g$ and $\mathcal{K}_l$ are unnormalized kernel families. Let us consider the normalized Gaussian family, $\mathcal{K}^n_g = \{(\sigma/\pi)^{d/2}\exp(-\sigma\|x-y\|^2_2),\ x,y \in$

$\mathbb{R}^d : \sigma \in [\sigma_0, \infty)\}$. It can be shown that for any k_σ, $k_\tau \in \mathcal{K}_g^n$, $0 < \sigma < \tau < \infty$, we have $\gamma_{k_\sigma}(\mathbb{P}, \mathbb{Q}) \ge \gamma_{k_\tau}(\mathbb{P}, \mathbb{Q})$, which means, $\gamma(\mathbb{P}, \mathbb{Q}) = \gamma_{\sigma_0}(\mathbb{P}, \mathbb{Q})$. Therefore, the generalized MMD reduces to a single kernel MMD. A similar result also holds for the normalized inverse-quadratic kernel family, $\{\sqrt{2\sigma^2/\pi}(\sigma^2 + \|x - y\|_2^2)^{-1}$, $x, y \in \mathbb{R} : \sigma \in [\sigma_0, \infty)\}$. These examples show that the generalized MMD definition is usually not very useful if $\mathcal{K}$ is a normalized kernel family. In addition, σ_0 should be chosen beforehand, which is equivalent to heuristically setting the kernel parameter in γ_k. Note that σ_0 cannot be zero because in the limiting case of $\sigma \to 0$, the kernels approach a Dirac distribution, which means the limiting kernel is not bounded and so the definition of MMD does not hold (see Proposition 3.2). Therefore, we restrict ourselves to unnormalized kernel families to render the definition of generalized MMD in (6.1) useful.

6.2 Estimation of γ: Consistency and Rate of Convergence

Given $\mathbb{P}$ and $\mathbb{Q}$, let us consider the computation of $\gamma(\mathbb{P}, \mathbb{Q})$, i.e.,

$$\gamma^2(\mathbb{P}, \mathbb{Q}) = \sup_{k \in \mathcal{K}} \gamma_k^2(\mathbb{P}, \mathbb{Q}) \overset{(3.6)}{=} \sup_{k \in \mathcal{K}} \iint_{\mathcal{X}} k(x, y)\, d(\mathbb{P} - \mathbb{Q})(x)\, d(\mathbb{P} - \mathbb{Q})(y).$$

Since the computation of $\gamma(\mathbb{P}, \mathbb{Q})$ is not straightforward as it involves taking supremum over $k \in \mathcal{K}$, similar to the case with $\gamma_k(\mathbb{P}, \mathbb{Q})$, we consider the approximation of $\gamma(\mathbb{P}, \mathbb{Q})$ as $\gamma(\mathbb{P}_m, \mathbb{Q}_n)$ and hope that $\gamma(\mathbb{P}_m, \mathbb{Q}_n)$ is strongly consistent. The strong consistency of $\gamma(\mathbb{P}_m, \mathbb{Q}_n)$ is also required in statistical applications where $\mathbb{P}$ and $\mathbb{Q}$ are known only through i.i.d. samples $\{X_j^{(1)}\}_{j=1}^m$ and $\{X_j^{(2)}\}_{j=1}^n$ respectively. For $\mathcal{K} = \{k\}$, where k is measurable and bounded, [37] has shown that $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ is a $\sqrt{mn/(m+n)}$-consistent estimator of $\gamma_k(\mathbb{P}, \mathbb{Q})$—also see Corollary 5.12(ii). Under certain conditions on $\mathcal{K}$, in the following we establish the strong consistency of $\gamma(\mathbb{P}_m, \mathbb{Q}_n)$. Before that, we consider the computation of $\gamma(\mathbb{P}_m, \mathbb{Q}_n)$, i.e.,

$$\gamma^2(\mathbb{P}_m, \mathbb{Q}_n) = \sup_{k \in \mathcal{K}} \left[\sum_{l,j=1}^{m} \frac{k(X_l^{(1)}, X_j^{(1)})}{m^2} + \sum_{l,j=1}^{n} \frac{k(X_l^{(2)}, X_j^{(2)})}{n^2} - 2 \sum_{l,j=1}^{m,n} \frac{k(X_l^{(1)}, X_j^{(2)})}{mn} \right].$$

In the following, we present examples on the computation of $\gamma(\mathbb{P}_m,\mathbb{Q}_n)$ for certain choices of $\mathcal{K}$.

Example 6.1. *Suppose $\mathcal{K}=\mathcal{K}_g$. Then $\gamma(\mathbb{P}_m,\mathbb{Q}_n)$ can be written as*

$$\begin{aligned}\gamma^2(\mathbb{P}_m,\mathbb{Q}_n) = \sup_{\sigma\in\mathbb{R}_+}\Bigg[&\frac{1}{m^2}\sum_{l,j=1}^{m} e^{-\sigma\left\|X_l^{(1)}-X_j^{(1)}\right\|_2^2}+\frac{1}{n^2}\sum_{l,j=1}^{n} e^{-\sigma\left\|X_l^{(2)}-X_j^{(2)}\right\|_2^2}\\ &-\frac{2}{mn}\sum_{l=1}^{m}\sum_{j=1}^{n} e^{-\sigma\left\|X_l^{(1)}-X_j^{(2)}\right\|_2^2}\Bigg],\end{aligned}$$

which is the maximization of a non-convex objective over the constraint set, $\Sigma := \{\sigma\,:\,\sigma\ge 0\}$—similar is the case for $\mathcal{K}=\mathcal{K}_l$ and $\mathcal{K}=\mathcal{K}_{rbf}$. Therefore, $\gamma(\mathbb{P}_m,\mathbb{Q}_n)$ cannot be computed in a simple closed form unlike $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ (in this case $\Sigma=\{\sigma_0\}$ for some $\sigma_0\ge 0$).

Example 6.2. *Suppose $\mathcal{K}=\mathcal{K}_{con}$. Then $\gamma(\mathbb{P}_m,\mathbb{Q}_n)$ becomes*

$$\begin{aligned}\gamma^2(\mathbb{P}_m,\mathbb{Q}_n) &= \sup\left\{\sum_{j=1}^{s}\lambda_j\gamma^2_{k_j}(\mathbb{P}_m,\mathbb{Q}_n)\,:\,\sum_{j=1}^{s}\lambda_j=1,\,\lambda_j\ge 0,\,\forall\, j\right\}\\ &= \max_{j\in\{1,\dots,s\}}\gamma^2_{k_j}(\mathbb{P}_m,\mathbb{Q}_n).\end{aligned}$$

Example 6.3. *Suppose $\mathcal{K}=\mathcal{K}_{lin}$. Then*

$$\gamma^2(\mathbb{P}_m,\mathbb{Q}_n) = \sup\left\{\sum_{j=1}^{s}\lambda_j\gamma^2_{k_j}(\mathbb{P}_m,\mathbb{Q}_n)\,:\,\sum_{j=1}^{s}\lambda_j k_j \text{ is pd},\,\sum_{j=1}^{s}\lambda_j=1\right\},$$

which is a semidefinite program as it is the maximization of a linear objective over a pd cone.

Now, we analyze the convergence of $\gamma(\mathbb{P}_m,\mathbb{Q}_n)$, for which we begin with the following definition.

Definition 6.4 (Rademacher chaos complexity)**.** *Let $\mathcal{G}$ be a class of functions on $\mathcal{X}\times\mathcal{X}$ and $\{\varrho_j\}_{j=1}^n$ be independent Rademacher random variables, i.e.,* $\Pr(\varrho_j = 1)=\Pr(\varrho_j=-1)=\frac{1}{2}$. *The homogeneous Rademacher chaos process of order two with respect to $\{\varrho_j\}_{j=1}^n$ is defined as*

$$\left\{\frac{1}{n}\sum_{l<j}^{n}\varrho_l\varrho_j g(X_l,X_j)\,:\,g\in\mathcal{G}\right\}$$

for some $\{X_j\}_{j=1}^n \subset \mathcal{X}$*. The Rademacher chaos complexity over* $\mathcal{G}$ *is defined as*

$$U_n(\mathcal{G}; \{X_j\}_{j=1}^n) := \mathbb{E} \sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum_{l<j}^{n} \varrho_l \varrho_j g(X_l, X_j) \right|.$$

The following result provides a probabilistic bound for the deviation of $\gamma(\mathbb{P}_m, \mathbb{Q}_n)$ from $\gamma(\mathbb{P}, \mathbb{Q})$ in terms of the Rademacher chaos complexity.

Theorem 6.5 (Consistency of $\gamma(\mathbb{P}_m, \mathbb{Q}_n)$)**.** *Let every* $k \in \mathcal{K}$ *be measurable and bounded with* $\nu := \sup_{k\in\mathcal{K}, x\in\mathcal{X}} \sqrt{k(x,x)} < \infty$*. Then, with probability at least* $1-\delta$*,*

$$\begin{aligned}|\gamma(\mathbb{P}_m, \mathbb{Q}_n) - \gamma(\mathbb{P}, \mathbb{Q})| \le \sqrt{\frac{8U_m(\mathcal{K}; \{X_j^{(1)}\}_{j=1}^m)}{m}} + \sqrt{\frac{8U_n(\mathcal{K}; \{X_j^{(2)}\}_{j=1}^n)}{n}} \\ + \left(2\nu + 3\nu\sqrt{2\log\frac{4}{\delta}}\right)\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right). \qquad (6.2)\end{aligned}$$

Proof. Define $\mathbb{P}f := \int_{\mathcal{X}} f \, d\mathbb{P}$. Since

$$\gamma(\mathbb{P}, \mathbb{Q}) = \sup_{k\in\mathcal{K}} \gamma_k(\mathbb{P}, \mathbb{Q}) = \sup_{k\in\mathcal{K}} \sup_{\|f\|_{\mathcal{H}_k}\le 1} |\mathbb{P}f - \mathbb{Q}f| = \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}),$$

where $\mathcal{F} = \cup_{k\in\mathcal{K}}\{f : \|f\|_{\mathcal{H}_k} \le 1\}$, the result follows by invoking Theorem 5.11, where we use

$$\begin{aligned}R_m(\mathcal{F}; \{X_j^{(1)}\}_{j=1}^m) &= \mathbb{E} \sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{j=1}^m \varrho_j f(X_j^{(1)}) \right| \\ &= \mathbb{E} \sup_{k\in\mathcal{K}} \sup_{\|f\|_{\mathcal{H}_k}\le 1} \left| \frac{1}{m} \sum_{j=1}^m \varrho_j \left\langle f, k(\cdot, X_j^{(1)}) \right\rangle_{\mathcal{H}_k} \right| \\ &= \mathbb{E} \sup_{k\in\mathcal{K}} \sup_{\|f\|_{\mathcal{H}_k}\le 1} \left| \frac{1}{m} \left\langle f, \sum_{j=1}^m \varrho_j k(\cdot, X_j^{(1)}) \right\rangle_{\mathcal{H}_k} \right| \\ &= \mathbb{E} \sup_{k\in\mathcal{K}} \left\| \frac{1}{m} \sum_{j=1}^m \varrho_j k(\cdot, X_j^{(1)}) \right\|_{\mathcal{H}_k} \\ &= \frac{1}{m} \mathbb{E} \sup_{k\in\mathcal{K}} \sqrt{\left| \sum_{l,j=1}^m \varrho_l \varrho_j k(X_l^{(1)}, X_j^{(1)}) \right|} \\ &\le \sqrt{\frac{2}{m}} \mathbb{E} \sqrt{\sup_{k\in\mathcal{K}} \left| \frac{1}{m} \sum_{l<j}^m \varrho_l \varrho_j k(X_l^{(1)}, X_j^{(1)}) \right|} + \sqrt{\frac{\nu^2}{m}}\end{aligned}$$

$$\overset{(\star)}{\leq} \sqrt{\frac{2}{m}}\sqrt{\mathbb{E}\sup_{k\in\mathcal{K}}\left|\frac{1}{m}\sum_{l<j}^{m}\varrho_l\varrho_j k(X_l^{(1)},X_j^{(1)})\right|}+\sqrt{\frac{\nu^2}{m}}$$
$$= \sqrt{\frac{2U_m(\mathcal{K};\{X_j^{(1)}\}_{j=1}^m)}{m}}+\sqrt{\frac{\nu^2}{m}}.$$

Here, we have invoked Jensen's inequality [26, p. 109] in ($\star$). □

From (6.2), it is clear that if $U_m(\mathcal{K};\{X_j^{(1)}\}_{j=1}^m) = O_{\mathbb{P}}(1)$ and $U_n(\mathcal{K};\{X_j^{(2)}\}_{j=1}^n) = O_{\mathbb{Q}}(1)$, then $|\gamma(\mathbb{P}_m,\mathbb{Q}_n)-\gamma(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(\sqrt{(m+n)/mn})$, which by the Borel-Cantelli lemma [23, Theorem 8.3.4] yields $\gamma(\mathbb{P}_m,\mathbb{Q}_n)\overset{a.s.}{\to}\gamma(\mathbb{P},\mathbb{Q})$. The following result provides a bound on $U_m(\mathcal{K};\{X_j^{(1)}\}_{j=1}^m)$ in terms of the entropy integral.

Lemma 6.6 (Entropy bound)**.** *For any $\mathcal{K}$ as in Theorem 6.5 with $0\in\mathcal{K}$, there exists a universal constant C such that*

$$U_m(\mathcal{K};\{X_j^{(1)}\}_{j=1}^m)\leq C\int_0^{\nu^2}\log\mathcal{N}(\mathcal{K},D,\epsilon)\,d\epsilon, \tag{6.3}$$

where $D(k_1,k_2)=\frac{1}{m}\left(\sum_{l<j}^m(k_1(X_l^{(1)},X_j^{(1)})-k_2(X_l^{(1)},X_j^{(1)}))^2\right)^{\frac{1}{2}}$. $\mathcal{N}(\mathcal{K},D,\epsilon)$ represents the ϵ-covering number of $\mathcal{K}$ with respect to the metric D.

Proof. From [3, Proposition 2.2, Proposition 2.6] (also see [18, Corollary 5.1.8]), we have that there exists a universal constant $C<\infty$ such that

$$U_m(\mathcal{K};\{X_j^{(1)}\}_{j=1}^m)\leq C\int_0^{\nu^2}\log\mathcal{N}(\mathcal{K},D,\epsilon)\,d\epsilon,$$

where

$$\begin{aligned}D^2(k_1,k_2) &= \mathbb{E}\left(\frac{1}{m}\sum_{l<j}^{n}\varrho_l\varrho_j h(X_l^{(1)},X_j^{(1)})\right)^2\\ &= \frac{1}{m^2}\mathbb{E}\left(\sum_{l<j,r<s}^{m}\varrho_l\varrho_j\varrho_r\varrho_s h(X_l^{(1)},X_j^{(1)})h(X_r^{(1)},X_s^{(1)})\right)\\ &= \frac{1}{m^2}\sum_{l<j}^{m}h^2(X_l^{(1)},X_j^{(1)}),\end{aligned}$$

and $h(X_l^{(1)},X_j^{(1)})=k_1(X_l^{(1)},X_j^{(1)})-k_2(X_l^{(1)},X_j^{(1)})$. □

Assuming $\mathcal{K}$ to be a Vapnik-Červonenkis (VC)-subgraph class, the following result, as a corollary to Lemma 6.6 provides an estimate of $U_m(\mathcal{K}; \{X_j^{(1)}\}_{j=1}^m)$. Before presenting the result, we first provide the definition of a VC-subgraph class (see [86, Section 2.6.2]).

Definition 6.7 (VC-subgraph class)**.** *Let $\mathcal{C}$ be a collection of subsets of a set $\mathcal{X}$ and let $\{x_1, \ldots, x_n\}$ be an arbitrary set of n points. The VC-index, $V(\mathcal{C})$ of the class $\mathcal{C}$ is defined as*[20]

$$V(\mathcal{C}) = \inf \left\{ n \, : \, \max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) < 2^n \right\},$$

where

$$\Delta_n(\mathcal{C}, x_1, \ldots, x_n) = \left| \{ C \cap \{x_1, \ldots, x_n\} \, : \, C \in \mathcal{C} \} \right|.$$

The subgraph of a function $g : \mathcal{X} \times \mathbb{R}$ is the subset of $\mathcal{X} \times \mathbb{R}$ given by $\{(x,t) : t < g(x)\}$. A collection $\mathcal{G}$ of measurable functions on a sample space is called a VC-subgraph class, if the collection of all subgraphs of the functions in $\mathcal{G}$ forms a VC-class of sets (in $\mathcal{X} \times \mathbb{R}$).[21]

Corollary 6.8 (Rademacher chaos complexity for VC-subgraph)**.** *Suppose $\mathcal{K}$ is a VC-subgraph class with $V(\mathcal{K})$ being the VC-index. Assume $\mathcal{K}$ satisfies the conditions in Theorem 6.5 and $0 \in \mathcal{K}$. Then*

$$U_m(\mathcal{K}; \{X_j^{(1)}\}_{j=1}^m) \le C\nu^2 \log \left(C_1 V(\mathcal{K}) \left(16 e^{1+8\nu^{-1/2}} \right)^{V(\mathcal{K})} \right), \tag{6.4}$$

for some universal constants C and C_1.

Proof. The result follows by bounding the uniform covering number of the VC-subgraph class, $\mathcal{K}$. By [86, Theorem 2.6], we have

$$\mathcal{N}(\mathcal{K}, D, \epsilon) \le C_1 V(\mathcal{K}) (16\nu^2 \epsilon^{-2} e)^{V(\mathcal{K})}.$$

[20] An arbitrary set of n points $A_n := \{x_1, \ldots, x_n\}$ possesses 2^n subsets. $\mathcal{C}$ is said to *pick out* a certain subset, B from A_n if $B = A_n \cap C$ for $C \in \mathcal{C}$. $\mathcal{C}$ is said to *shatter* A_n if each of its 2^n subsets can be picked out in this manner. The VC-index of $\mathcal{C}$ is the smallest n for which no set of size n is shattered by $\mathcal{C}$.

[21] The VC-index (also called the VC-dimension) of a VC-subgraph class, $\mathcal{G}$ is the same as the *pseudo-dimension* of $\mathcal{G}$. See [2, Definition 11.1] for details.

Therefore, from (6.3), we have

$$\begin{aligned} U_m(\mathcal{K};\{X^{(1)}_j\}_{j=1}^m) &\le C\int_0^{\nu^2} \log\mathcal{N}(\mathcal{K},D,\epsilon)\,d\epsilon \\ &\le \nu^2\log\left(C_1V(\mathcal{K})(16e)^{V(\mathcal{K})}\right)+4V(\mathcal{K})C\int_0^{\nu^2}\log\left(\frac{\sqrt{\nu}}{\sqrt{\epsilon}}\right)d\epsilon \\ &\overset{(\star)}{\le} \nu^2\log\left(C_1V(\mathcal{K})(16e)^{V(\mathcal{K})}\right)+8V(\mathcal{K})C\nu^{3/2} \\ &= C\nu^2\log\left(C_1V(\mathcal{K})\left(16e^{1+8\nu^{-1/2}}\right)^{V(\mathcal{K})}\right), \end{aligned}$$

where we have used $\log x \le x$ in $(\star)$. □

Using (6.4) in (6.2), we have $|\gamma(\mathbb{P}_m,\mathbb{Q}_n)-\gamma(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(\sqrt{(m+n)/mn})$ and by the Borel-Cantelli lemma [23, Theorem 8.3.4], $|\gamma(\mathbb{P}_m,\mathbb{Q}_n)-\gamma(\mathbb{P},\mathbb{Q})| \xrightarrow{a.s.} 0$. Now, the question reduces to which of the kernel classes, $\mathcal{K}$ have $V(\mathcal{K})<\infty$. [93, Lemma 12] showed that $V(\mathcal{K}_g)=1$ (also see [94]) and $U_m(\mathcal{K}_{rbf};\{X^{(1)}_j\}_{j=1}^m) \le C_2U_m(\mathcal{K}_g;\{X^{(1)}_j\}_{j=1}^m)$, where $C_2<\infty$. It can be shown that $V(\mathcal{K}_\psi)=1$ and $V(\mathcal{K}_l)=1$. [74, Lemma 7] has shown that $V(\mathcal{K}_{\text{con}}) \le V(\mathcal{K}_{\text{lin}}) \le l$. Since all these classes satisfy the conditions of Theorem 6.5 and Corollary 6.8, they provide consistent estimates of $\gamma(\mathbb{P},\mathbb{Q})$ for any $\mathbb{P},\mathbb{Q}\in M^1_+(\mathscr{X})$. Examples of kernels on $\mathbb{R}^d$ that are covered by these classes include the Gaussian, Laplacian, inverse multiquadratics, Matérn class etc.

6.3 Experiments

In this section, we present a benchmark experiment that illustrates the generalized MMD proposed in Section 6.1 is preferred above the single kernel MMD where the kernel parameter is set heuristically. The experimental setup is as follows.

Let $p=N(0,\sigma_p^2)$, a normal distribution in $\mathbb{R}$ with zero mean and variance, σ_p^2. Let q be the perturbed version of p, given as $q(x)=p(x)(1+\sin\nu x)$. Here p and q are the densities associated with $\mathbb{P}$ and $\mathbb{Q}$ respectively. It is easy to see that q differs from p at increasing frequencies with increasing ν. Let $k(x,y)=\exp(-(x-y)^2/\sigma)$. Now, the goal is that given random samples drawn i.i.d. from

$\mathbb{P}$ and $\mathbb{Q}$ (with ν fixed), we would like to test $H_0 : \mathbb{P} = \mathbb{Q}$ vs. $H_1 : \mathbb{P} \neq \mathbb{Q}$. The idea is that as ν increases, it will be harder to distinguish between $\mathbb{P}$ and $\mathbb{Q}$ for a fixed sample size. Therefore, using this setup we can verify whether the adaptive bandwidth selection achieved by γ (as the test statistic) helps to distinguish between $\mathbb{P}$ and $\mathbb{Q}$ at higher ν compared to γ_k with a heuristic σ. To this end, using $\gamma(\mathbb{P}_m, \mathbb{Q}_n)$ and $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ (with various σ) as test statistics T_{mn}, we design a test that returns H_0 if $T_{mn} \leq c_{mn}$, and H_1 otherwise. The problem therefore reduces to finding c_{mn}. c_{mn} is determined as the $(1 - \alpha)$ quantile of the asymptotic distribution of T_{mn} under H_0, which therefore fixes the type-I error (the probability of rejecting H_0 when it is true) to α. The consistency of this test under γ_k (for any fixed σ) is proved in [37]. A similar result can be shown for γ under some conditions on $\mathcal{K}$. We skip the details here.

In our experiments, we set $m = n = 1000$, $\sigma_p^2 = 10$ and draw two sets of independent random samples from $\mathbb{Q}$. The distribution of T_{mn} is estimated by bootstrapping on these samples (250 bootstrap iterations are performed) and the associated 95^{th} quantile (we choose $\alpha = 0.05$) is computed. Since the performance of the test is judged by its type-II error (the probability of accepting H_0 when H_1 is true), we draw a random sample, one each from $\mathbb{P}$ and $\mathbb{Q}$ and test whether $\mathbb{P} = \mathbb{Q}$. This process is repeated 300 times, and estimates of type-I and type-II errors are obtained for both γ and γ_k. 14 different values for σ are considered on a logarithmic scale of base 2 with exponents $(-3, -2, -1, 0, 1, \frac{3}{2}, 2, \frac{5}{2}, 3, \frac{7}{2}, 4, 5, 6)$ along with the median distance between samples as one more choice. 5 different choices for ν are considered: $(\frac{1}{2}, \frac{3}{4}, 1, \frac{5}{4}, \frac{3}{2})$.

Figure 6.1(a) shows the estimated type-I and type-II errors using γ as the test statistic for varying ν. Note that the type-I error is close to its design value of 5%, while the type-II error is zero for all ν, which means γ distinguishes between $\mathbb{P}$ and $\mathbb{Q}$ for all perturbations. Figures 6.1(b,c) show the estimates of type-I and type-II errors using γ_k as the test statistic for different σ and ν. Figure 6.1(d) shows the box plot for $\log \sigma$, grouped by ν, where σ is the bandwidth selected by γ. Figure 6.1(e) shows the box plot of the median distance between points (which is also a choice for σ), grouped by ν. From Figures 6.1(c) and (e), it is easy to see that the median heuristic exhibits high type-II error for $\nu = \frac{3}{2}$, while γ exhibits zero type-II error (from Figure 6.1(a)). Figure 6.1(c) also shows that heuristic choices of σ can result in high type-II errors. It is intuitive to note that as ν increases, (which means the characteristic function of $\mathbb{Q}$ differs from that of $\mathbb{P}$ at higher frequencies), a smaller σ is needed to detect these changes. The advantage of using γ is that it selects σ in a distribution-dependent fashion and its behavior in the box plot shown in Figure 6.1(d) matches with the previously mentioned intuition about the behavior of σ with respect to ν. These results demonstrate the validity of using γ as a distance measure in applications.

Bibliographic Notes

This chapter is based on joint work with Kenji Fukumizu, Arthur Gretton, Gert Lanckriet and Bernhard Schölkopf, which appeared in [75]. The dissertation author was the primary investigator and author of this paper.

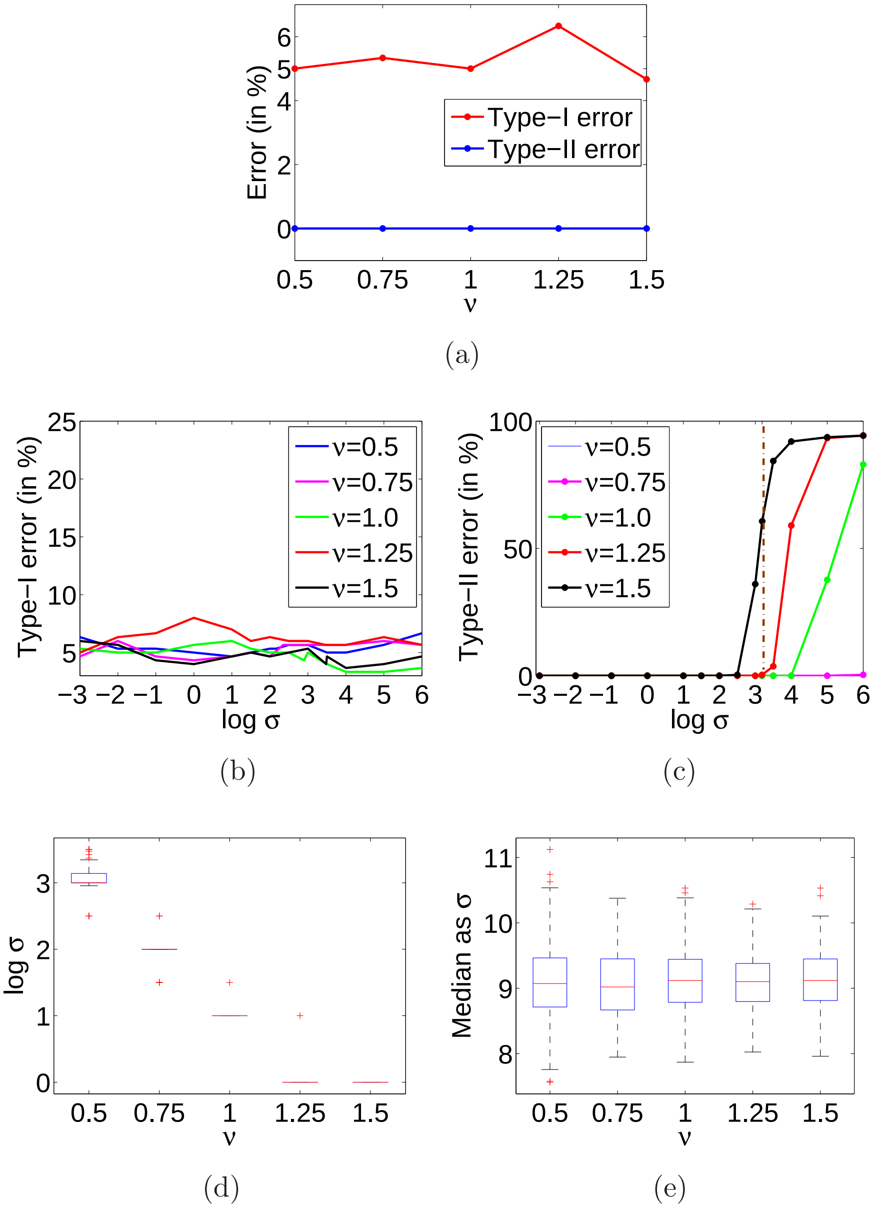


Figure 6.1: (a) Type-I and Type-II errors (in %) for γ for varying ν. (b,c) Type-I and type-II error (in %) for γ_k (with different σ) for varying ν. The dotted line in (c) corresponds to the median heuristic, which shows that its associated type-II error is very large at large ν. (d) Box plot of $\log \sigma$ grouped by ν, where σ is selected by γ. (e) Box plot of the median distance between points (which is also a choice for σ), grouped by ν. Refer to Section 6.3 for details.

7 Banach Space Embedding of Probability Measures

So far, in Chapters 3–6, we have studied the embedding $\mathbb{P} \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$ and the associated pseudometric, γ_k on $M^1_+(\mathcal{X})$, where k is a reproducing kernel associated with an RKHS, $\mathcal{H}$. The goal of this chapter is to generalize this notion of RKHS embedding of probability measures to Banach spaces, in particular reproducing kernel Banach spaces (RKBSs) [95]. This is primarily based on two different motivations. Firstly, Banach spaces possess much richer geometric structure than Hilbert spaces in sense that any two Hilbert spaces over $\mathbb{C}$ of the same dimension are isometrically isomorphic while this is not the case with Banach spaces, e.g., for $p \neq q \in [1, \infty]$, $L^p[0, 1]$ and $L^q[0, 1]$ are not isomorphic. Therefore, "richer" distance measures between probabilities could be obtained by embedding them into more general spaces like an RKBS. Secondly, since we have shown the connection between binary classification and the RKHS embedding of probability measures (see Section 5.2.2) and since binary classification algorithms have been studied in Banach spaces [20, 88, 95], one can obtain an alternate view of classification in Banach spaces through the notion of probability embeddings in Banach spaces. RKBSs were recently studied by Zhang et al. [95] in the context of developing learning algorithms in Banach spaces, wherein many RKHS based algorithms like regularization networks, support vector machines, kernel principal component analysis, etc., were extended to RKBS. In this chapter, we investigate how the notion of RKHS embedding of probability measures extends to an RKBS and the similarities/differences in the properties of an RKBS embedding compared to its RKHS counterpart, along with its advantages/disadvantages.

The chapter is organized as follows. In Section 7.1, following [95], we provide preliminaries of RKBS. In Section 7.2, we *first* derive an RKBS embedding of $\mathbb{P}$ into $\mathcal{B}'$ as

$$\mathbb{P} \mapsto \int_{\mathcal{X}} K(\cdot, x)\, d\mathbb{P}(x), \tag{7.1}$$

where $\mathcal{B}$ is a *uniformly Fréchet differentiable* and *uniformly convex* RKBS with K as its reproducing kernel (r.k.) and $\mathcal{B}'$ is the topological dual of $\mathcal{B}$. Note that (7.1) is similar to (3.1), but more general than (3.1) as K in (7.1) need not have to be positive definite, in fact, not even symmetric (see Section 7.1). *Second*, we characterize the injectivity of (7.1) in Section 7.2.1 wherein we show that the characterizations obtained for the injectivity of (7.1) are very similar to those obtained for (3.1) and coincide with the latter when $\mathcal{B}$ is an RKHS. Based on (7.1), we define

$$\gamma_K(\mathbb{P}, \mathbb{Q}) := \left\| \int_{\mathcal{X}} K(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot, x)\, \mathbb{Q}(x) \right\|_{\mathcal{B}'},$$

which is a pseudo-metric on $M_+^1(\mathcal{X})$. *Third*, in Section 7.2.2, we consider the empirical estimation of $\gamma_K(\mathbb{P}, \mathbb{Q})$ based on finite random samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$ and study its consistency and the rate of convergence. This is useful in applications like two-sample tests where different $\mathbb{P}$ and $\mathbb{Q}$ are to be distinguished based on the finite samples drawn from them and it is important that the estimator is consistent for the test to be meaningful. We show that the consistency and the rate of convergence of the estimator depend on the *Rademacher type* of $\mathcal{B}'$. This result coincides with the one obtained for γ_k when $\mathcal{B}$ is an RKHS (see Corollary 5.12(ii)).

The above mentioned results, while similar to results obtained for RKHS embeddings, are significantly more general, as they apply RKBS spaces, which subsume RKHSs. We can therefore expect to obtain "richer" metrics γ_K than when being restricted to RKHSs. On the other hand, one disadvantage of the RKBS framework is that $\gamma_K(\mathbb{P}, \mathbb{Q})$ cannot be computed in a closed form unlike γ_k (see Section 7.2.3). This could seriously limit the practical impact of these results, as is the case with the RKBS based learning algorithms derived by [95], which are not straightforward to implement. The proposed theory of RKBS embeddings of probability measures, however, *does* have a practical impact as closed form solutions can be obtained in some cases. In Section 7.3, we provide concrete examples of s.i.p. RKBS for which the RKBS embeddings and the corresponding $\gamma_K(\mathbb{P},\mathbb{Q})$ can be obtained in a closed form.

7.1 Preliminaries: Reproducing Kernel Banach Spaces

In this section, we briefly review the theory of reproducing kernel Banach spaces, which was recently studied by Zhang et al. [95] in the context of learning in Banach spaces. Let $\mathcal{X}$ be a prescribed input space.

Definition 7.1 (Reproducing kernel Banach space)**.** *An RKBS $\mathcal{B}$ on $\mathcal{X}$ is a reflexive Banach space of functions on $\mathcal{X}$ such that its topological dual $\mathcal{B}'$ is isometric to a Banach space of functions on $\mathcal{X}$ and the point evaluations are continuous linear functionals on both $\mathcal{B}$ and $\mathcal{B}'$.*

Note that if $\mathcal{B}$ is a Hilbert space, then the above definition of RKBS coincides with that of an RKHS. Let $(\cdot,\cdot)_{\mathcal{B}}$ be a bilinear form on $\mathcal{B}\times\mathcal{B}'$ wherein $(f,g^*)_{\mathcal{B}} := g^*(f)$, $f\in\mathcal{B}$, $g^*\in\mathcal{B}'$. Theorem 2 in [95] shows that if $\mathcal{B}$ is an RKBS on $\mathcal{X}$, then there exists a unique function $K:\mathcal{X}\times\mathcal{X}\to\mathbb{C}$ called the reproducing kernel (r.k.) of $\mathcal{B}$, such that the following hold:

(a_1) $K(x,\cdot)\in\mathcal{B}$, $K(\cdot,x)\in\mathcal{B}'$, $x\in\mathcal{X}$,

(a_2) $f(x)=(f,K(\cdot,x))_{\mathcal{B}}$, $f^*(x)=(K(x,\cdot),f^*)_{\mathcal{B}}$, $f\in\mathcal{B}$, $f^*\in\mathcal{B}'$, $x\in\mathcal{X}$.

Note that K satisfies $K(x,y)=(K(x,\cdot),K(\cdot,y))_{\mathcal{B}}$. When $\mathcal{B}$ is an RKHS, K is indeed the r.k. in the usual sense. Though an RKBS has exactly one r.k., different RKBSs may have the same r.k. (see Example 7.17 in Section 7.3) unlike an RKHS, where no two RKHSs can have the same r.k (by the Moore-Aronszajn Theorem). Due to the lack of inner product in $\mathcal{B}$ (unlike in an RKHS), Zhang et al. have shown that the r.k. for a general RKBS can be any arbitrary function on $\mathcal{X}\times\mathcal{X}$. Therefore, to have a substitute for inner products in the Banach space setting, they considered RKBS $\mathcal{B}$ that are uniformly Fréchet differentiable and uniformly convex (referred to as s.i.p. RKBS) as it allows Hilbert space arguments to be carried over to $\mathcal{B}$—most importantly, an analogue to the Riesz representation theorem holds (see Theorem 7.3)—through the notion of *semi-inner-product* (s.i.p.) introduced by [50]. In the following, we first present results related to general s.i.p. spaces (Banach spaces that are uniformly Fréchet differentiable and uniformly convex) and then consider s.i.p. RKBS.

Definition 7.2 (S.i.p. space)**.** *A Banach space $\mathcal{B}$ is said to be uniformly Fréchet differentiable if for all $f, g \in \mathcal{B}$,*

$$\lim_{t\in\mathbb{R}, t\to 0} \frac{\|f+tg\|_{\mathcal{B}} - \|f\|_{\mathcal{B}}}{t}$$

exists and the limit is approached uniformly for f, g in the unit sphere of $\mathcal{B}$. $\mathcal{B}$ is said to be uniformly convex if for all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\|f+g\|_{\mathcal{B}} \le 2-\delta \textit{ for all } f, g \in \mathcal{B} \textit{ with } \|f\|_{\mathcal{B}} = \|g\|_{\mathcal{B}} = 1 \textit{ and } \|f-g\|_{\mathcal{B}} \ge \epsilon.$$

$\mathcal{B}$ is called an s.i.p. space if it is both uniformly Fréchet differentiable and uniformly convex.

Note that uniform Fréchet differentiability and uniform convexity are properties of the norm associated with $\mathcal{B}$. Giles [34, Theorem 3] has shown that if $\mathcal{B}$ is an s.i.p. space, then there exists a unique function $[\cdot,\cdot]_{\mathcal{B}} : \mathcal{B}\times\mathcal{B} \to \mathbb{C}$, called the semi-inner-product such that for all $f, g, h \in \mathcal{B}$ and $\lambda \in \mathbb{C}$:

(a_3) $[f+g, h]_{\mathcal{B}} = [f, h]_{\mathcal{B}} + [g, h]_{\mathcal{B}}$,

(a_4) $[\lambda f, g]_{\mathcal{B}} = \lambda[f, g]_{\mathcal{B}}$, $[f, \lambda g]_{\mathcal{B}} = \overline{\lambda}[f, g]_{\mathcal{B}}$,

(a_5) $[f, f]_{\mathcal{B}} =: \|f\|^2_{\mathcal{B}} > 0$ for $f \neq 0$,

(a_6) (Cauchy-Schwartz) $|[f, g]_{\mathcal{B}}|^2 \le \|f\|^2_{\mathcal{B}}\|g\|^2_{\mathcal{B}}$,

and

$$\lim_{t\in\mathbb{R}, t\to 0} \frac{\|f+tg\|_{\mathcal{B}} - \|f\|_{\mathcal{B}}}{t} = \frac{\mathrm{Re}([g, f]_{\mathcal{B}})}{\|f\|_{\mathcal{B}}},\ f, g \in \mathcal{B},\ f \neq 0,$$

where $\mathrm{Re}(\alpha)$ and $\overline{\alpha}$ represent the real part and complex conjugate of a complex number α. Note that semi-inner-products in general do not satisfy conjugate symmetry, $[f,g]_{\mathcal{B}} = \overline{[g,f]_{\mathcal{B}}}$ for all $f, g \in \mathcal{B}$ and therefore are not linear in the second argument, unless $\mathcal{B}$ is a Hilbert space, in which case the s.i.p. coincides with the inner product.

Suppose $\mathcal{B}$ is an s.i.p. space. Then for each $h \in \mathcal{B}$, $f \mapsto [f,h]_{\mathcal{B}}$ defines a continuous linear functional on $\mathcal{B}$, which can be identified with a unique element $h^* \in \mathcal{B}'$, called the *dual function* of h. By this definition of h^*, we have $h^*(f) = (f, h^*)_{\mathcal{B}} = [f,h]_{\mathcal{B}}$, $f, h \in \mathcal{B}$. Using the structure of s.i.p., Giles [34, Theorem 6] provided the following analogue in $\mathcal{B}$ to the Riesz representation theorem of Hilbert spaces.

Theorem 7.3 ([34]). *Suppose $\mathcal{B}$ is an s.i.p. space. Then*

(a_7) *(Riesz representation theorem) For each $g \in \mathcal{B}'$, there exists a unique $h \in \mathcal{B}$ such that $g = h^*$, i.e., $g(f) = [f,h]_{\mathcal{B}}$, $f \in \mathcal{B}$ and $\|g\|_{\mathcal{B}'} = \|h\|_{\mathcal{B}}$.*

(a_8) *$\mathcal{B}'$ is an s.i.p. space with respect to the s.i.p. defined by*

$$[h^*, f^*]_{\mathcal{B}'} := [f,h]_{\mathcal{B}},\ f, h \in \mathcal{B}$$

and $\|h^*\|_{\mathcal{B}'} := [h^*, h^*]_{\mathcal{B}'}^{1/2}$.

For more details on s.i.p. spaces, we refer the reader to [34]. A concrete example of an s.i.p. space is as follows, which will prove to be useful in Section 7.2. Let $(\mathcal{X}, \mathscr{A}, \mu)$ be a measure space and $\mathcal{B} := L^p(\mathcal{X}, \mu)$ for some $p \in (1, +\infty)$. It is an s.i.p. space with dual $\mathcal{B}' := L^q(\mathcal{X}, \mu)$. For each $f \in \mathcal{B}$, its dual element in $\mathcal{B}'$ is

$$f^* = \frac{\overline{f}|f|^{p-2}}{\|f\|_{L^p(\mathcal{X},\mu)}^{p-2}}.$$

Consequently, the semi-inner-product on $\mathcal{B}$ is

$$[f,g]_{\mathcal{B}} = g^*(f) = \frac{\int_{\mathcal{X}} f\overline{g}|g|^{p-2}\, d\mu}{\|g\|_{L^p(\mathcal{X},\mu)}^{p-2}}. \tag{7.2}$$

Having introduced s.i.p. spaces, we now discuss s.i.p. RKBS which was studied by [95]. Using the Riesz representation for s.i.p. spaces (see (a_7)), Theorem 9 in [95] shows that if $\mathcal{B}$ is an s.i.p. RKBS with K as its r.k., then there exists a unique s.i.p. kernel $G : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ such that:

(a_9) $G(x, \cdot) \in \mathcal{B}$ for all $x \in \mathcal{X}$, $K(\cdot, x) = (G(x, \cdot))^*$, $x \in \mathcal{X}$,

(a_{10}) $f(x) = [f, G(x, \cdot)]_{\mathcal{B}}$, $f^*(x) = [K(x, \cdot), f]_{\mathcal{B}}$ for all $f \in \mathcal{B}$, $x \in \mathcal{X}$.

It is clear that $G(x, y) = [G(x, \cdot), G(y, \cdot)]_{\mathcal{B}}$, $x, y \in \mathcal{X}$. Since s.i.p. in general do not satisfy conjugate symmetry, G need not have to be Hermitian nor pd [95, Section 4.3]. The r.k. K and the s.i.p. kernel G coincide when span$\{G(x, \cdot) : x \in \mathcal{X}\}$ is dense in $\mathcal{B}$, which is the case when $\mathcal{B}$ is an RKHS [95, Theorem 10]. This means when $\mathcal{B}$ is an RKHS, then the conditions (a_9) and (a_{10}) reduce to the well-known reproducing properties of an RKHS with the semi-inner-product reducing to an inner product.

7.2 RKBS Embedding of Probability Measures

In this section, we derive and analyze the RKBS embedding of probability measures, which generalize the theory of RKHS embeddings. First, we would like to remind the reader that the RKHS embedding in (3.1) can be derived by choosing $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\}$ in (3.2). Similar to the RKHS case, in Theorem 7.5, we show that the RKBS embeddings can be obtained by choosing $\mathcal{F} = \{f : \|f\|_{\mathcal{B}} \le 1\}$ in (3.2). Interestingly, though $\mathcal{B}$ does not have an inner product, it can be seen that the structure of semi-inner-product is sufficient enough to generate an embedding similar to (3.1). Before that, we need the following supplementary result (similar to Lemma 3.1), which will be useful to prove Theorem 7.5.

Lemma 7.4. *Let $\mathcal{B}$ be an s.i.p. RKBS defined on a measurable space $\mathcal{X}$ with G as the s.i.p. kernel and K as the reproducing kernel with both G and K being measurable and G bounded. Suppose μ be a finite signed measure on $\mathcal{X}$. Then, for any $f \in \mathcal{B}$, we have*

$$\int_{\mathcal{X}} f(x)\, d\mu(x) = \int_{\mathcal{X}} [K(\cdot, x), f^*]_{\mathcal{B}'}\, d\mu(x) = \left[\int_{\mathcal{X}} K(\cdot, x)\, d\mu(x), f^*\right]_{\mathcal{B}'}. \qquad (7.3)$$

Proof. The idea of the proof is similar to that of Lemma 3.1. Consider $T_\mu[f] = \int_{\mathcal{X}} f(x)\,d\mu(x)$. Since $\mathcal{B}$ is an s.i.p. RKBS, then by (a_{10}) there exists a unique G such that $f(x) = [f, G(x,\cdot)]_{\mathcal{B}} \overset{(a_8),(a_9)}{=} [K(\cdot,x), f^*]_{\mathcal{B}'}$. Therefore, we have

$$\begin{aligned} |T_\mu[f]| \overset{(a_{10})}{=} \left|\int_{\mathcal{X}} [f, G(x,\cdot)]_{\mathcal{B}}\,d\mu(x)\right| &\leq \int_{\mathcal{X}} |[f, G(x,\cdot)]_{\mathcal{B}}|\ d|\mu|(x) \\ &\overset{(a_6)}{\leq} \|f\|_{\mathcal{B}} \int_{\mathcal{X}} [G(x,\cdot), G(x,\cdot)]_{\mathcal{B}}^{1/2}\,d|\mu|(x) \\ &\overset{(a_{10})}{=} \|f\|_{\mathcal{B}} \int_{\mathcal{X}} \sqrt{G(x,x)}\,d|\mu|(x) < \infty, \end{aligned}$$

which means $T_\mu \in \mathcal{B}'$. By (a_7), there exists a unique $\lambda_\mu \in \mathcal{B}$ such that $T_\mu = \lambda_\mu^*$, i.e., $T_\mu[f] = [f, \lambda_\mu]_{\mathcal{B}}$, $f \in \mathcal{B}$. In other words,

$$\int_{\mathcal{X}} [f, G(x,\cdot)]_{\mathcal{B}}\,d\mu(x) = \int_{\mathcal{X}} f(x)\,d\mu(x) = T_\mu[f] = [f, \lambda_\mu]_{\mathcal{B}} \overset{(a_8)}{=} \left[\lambda_\mu^*, f^*\right]_{\mathcal{B}'}. \tag{7.4}$$

Choosing $f = K(y,\cdot) \in \mathcal{B}$ for some $y \in \mathcal{X}$ in (7.4) gives

$$\int_{\mathcal{X}} [K(y,\cdot), G(x,\cdot)]_{\mathcal{B}}\,d\mu(x) = \int_{\mathcal{X}} K(y,x)\,d\mu(x) = [K(y,\cdot), \lambda_\mu]_{\mathcal{B}} \overset{(a_{10})}{=} \lambda_\mu^*(y).$$

This means $\lambda_\mu^* = \int_{\mathcal{X}} K(\cdot,x)\,d\mu(x)$ and the result follows. □

Theorem 7.5. *Let $\mathcal{B}$ be an s.i.p. RKBS defined on a measurable space $\mathcal{X}$ with G as the s.i.p. kernel and K as the reproducing kernel with both G and K being measurable. Let $\mathcal{F} = \{f : \|f\|_{\mathcal{B}} \leq 1\}$ and G be bounded. Then*

$$\gamma_K(\mathbb{P},\mathbb{Q}) := \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \left\|\int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{Q}(x)\right\|_{\mathcal{B}'}. \tag{7.5}$$

Proof. Consider

$$\begin{aligned} \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) &= \sup_{\|f\|_{\mathcal{B}}\leq 1} \left|\int_{\mathcal{X}} f\,d\mathbb{P} - \int_{\mathcal{X}} f\,d\mathbb{Q}\right| \\ &\overset{(7.3)}{=} \sup_{\|f\|_{\mathcal{B}}\leq 1} \left|\left[\int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{Q}(x), f^*\right]_{\mathcal{B}'}\right| \\ &\overset{(a_8)}{=} \sup_{\|f^*\|_{\mathcal{B}'}\leq 1} \left|\left[\int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{Q}(x), f^*\right]_{\mathcal{B}'}\right| \\ &\overset{(a_6)}{=} \left\|\int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{Q}(x)\right\|_{\mathcal{B}'}, \end{aligned}$$

therefore proving the result. □

Based on Theorem 7.5, it is clear that $\mathbb{P}$ can be seen as being embedded into $\mathcal{B}'$ as

$$\mathbb{P} \mapsto \int_X K(\cdot,x)\,d\mathbb{P}(x) \tag{7.6}$$

and $\gamma_K(\mathbb{P},\mathbb{Q})$ is the distance between the embeddings of $\mathbb{P}$ and $\mathbb{Q}$. Therefore, we arrive at an embedding which looks similar to (3.1) and coincides with (3.1) when $\mathcal{B}$ is an RKHS.

Given these embeddings, two questions that need to be answered for these embeddings to be practically useful are: ($\star$) When is the embedding injective? and ($\star\star$) Can $\gamma_K(\mathbb{P},\mathbb{Q})$ in (7.5) be estimated consistently and computed efficiently from finite random samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$? We answered these questions in Chapters 3–5 when $\mathcal{B}$ is an RKHS. The significance of ($\star$) is that if (7.6) is injective, then such an embedding can be used to differentiate between different $\mathbb{P}$ and $\mathbb{Q}$, which can then be used in applications like two-sample test to differentiate between $\mathbb{P}$ and $\mathbb{Q}$ based on samples drawn i.i.d. from them if the answer to ($\star\star$) is affirmative. These questions are answered in the following sections.

7.2.1 When is (7.6) Injective?

The following result provides various characterizations for the injectivity of (7.6), which are very similar (but general) to those obtained for the injectivity of (3.1) and coincide with the latter when $\mathcal{B}$ is an RKHS. The proof technique is similar to that used to prove the results for RKHS embeddings.

Theorem 7.6. *Suppose $\mathcal{B}$ is an s.i.p. RKBS defined on a topological space $\mathcal{X}$ with K and G as its r.k. and s.i.p. kernel respectively. Then the following hold:*

(a) Let $\mathcal{X}$ be a Polish space that is also locally compact Hausdorff. Suppose G is bounded and $K(x,\cdot) \in C_0(\mathcal{X})$ for all $x \in \mathcal{X}$. Then (7.6) is injective if $\mathcal{B}$ is dense in $C_0(\mathcal{X})$.

(b) Suppose the conditions in (a) hold. Then (7.6) is injective if $\mathcal{B}$ is dense in $L^p(\mathcal{X},\mu)$ for any Borel probability measure μ on $\mathcal{X}$ and some $p \in [1,\infty)$.

Proof. *(a)* We first show that if G is bounded and $K(x,\cdot) \in C_0(\mathcal{X}), \forall\, x \in \mathcal{X}$, then

$\mathcal{B} \subset C_0(\mathcal{X})$. Since G is bounded, we have

$$|f(x)| = |[f, G(x,\cdot)]_{\mathcal{B}}| \le \|f\|_{\mathcal{B}}\sqrt{G(x,x)} \le \|f\|_{\mathcal{B}}\|G\|_\infty$$

for all $f \in \mathcal{B}$ and $x \in \mathcal{X}$, which means $\|f\|_\infty \le \|G\|_\infty\|f\|_{\mathcal{B}}, \forall\, f \in \mathcal{B}$. Here $\|G\|_\infty := \sup\{\sqrt{G(x,x)} : x \in \mathcal{X}\}$. This means $\mathrm{id} : \mathcal{B} \to \ell_\infty(\mathcal{X})$ is well-defined and $\|\mathrm{id} : \mathcal{B} \to \ell_\infty(\mathcal{X})\| \le \|G\|_\infty$, where $\ell_\infty(\mathcal{X})$ is the space of bounded functions on $\mathcal{X}$. Let us define $\mathcal{B}_{pre} := \mathrm{span}\{K(x,\cdot) : x \in \mathcal{X}\}$. Since $K(x,\cdot) \in C_0(\mathcal{X}), \forall\, x \in \mathcal{X}$, it is clear that $\mathcal{B}_{pre} \subset C_0(\mathcal{X})$. Theorem 2 in [95] shows that $\mathcal{B}_{pre}$ is dense in $\mathcal{B}$, which means for any $f \in \mathcal{B}$, there exists a sequence $\{f_n\} \subset \mathcal{B}_{pre}$ such that $\lim_{n\to\infty}\|f - f_n\|_{\mathcal{B}} = 0$ and the continuity of $\mathrm{id} : \mathcal{B} \to \ell_\infty(\mathcal{X})$ then yields $\lim_{n\to\infty}\|f - f_n\|_\infty = 0$. The completeness of $C_0(\mathcal{X})$ shows that $C_0(\mathcal{X})$ is a closed subspace of $\ell_\infty(\mathcal{X})$, and since $f_n \in C_0(\mathcal{X})$, $\forall\, n$, we can conclude that $f \in C_0(\mathcal{X})$. Therefore, the inclusion $\mathrm{id} : \mathcal{B} \to C_0(\mathcal{X})$ is well-defined and continuous.

We now show that if $\mathcal{B}$ is dense in $C_0(\mathcal{X})$, then (7.6) is injective. To show this, we first obtain an equivalent representation for the denseness of $\mathcal{B}$ in $C_0(\mathcal{X})$ and then show that if (7.6) is not injective, then $\mathcal{B}$ is not dense in $C_0(\mathcal{X})$, thereby proving the result. By the Hahn-Banach theorem (Theorem 4.7), $\mathcal{B}$ is dense in $C_0(\mathcal{X})$ if and only if $\mathcal{B}^\perp = \{\mu \in M_b(\mathcal{X}) : \forall\, f \in \mathcal{B}, \int_X f\, d\mu = 0\} = \{0\}$. Let us assume that $\mu \mapsto \int_X K(\cdot,x)\, d\mu(x)$, $\mu \in M_b(\mathcal{X})$ is not injective. This means there exists $\mu \in M_b(\mathcal{X})\backslash\{0\}$ such that $\int_X K(\cdot,x)\, d\mu(x) = 0$, which means $\int_X f(x)\, d\mu(x) = [\int_X K(\cdot,x)\, d\mu(x), f^*]_{\mathcal{B}'} = 0$ for any $f \in \mathcal{B}$, where we used (7.3). In other words, $\mathcal{B}^\perp \ne \{0\}$, which means $\mathcal{B}$ is not dense in $C_0(\mathcal{X})$. Therefore, if $\mathcal{B}$ is dense in $C_0(\mathcal{X})$, then $\mu \mapsto \int_X K(\cdot,x)\, d\mu(x)$, $\mu \in M_b(\mathcal{X})$ is injective, which means (7.6) is injective.

(b) Suppose the conditions in *(a)* hold. We claim that $\mathcal{B}$ is dense in $C_0(\mathcal{X})$ if and only if $\mathcal{B}$ is dense in $L^p(\mathcal{X},\mu)$ for all Borel probability measures μ on $\mathcal{X}$ and some $p \in [1,\infty)$. If this claim is true, then clearly the result in Theorem 7.6(b) follows. The proof of the claim is as follows, which is essentially based on [13, Theorem 1].

$(\Leftarrow)$ Suppose $\mathcal{B}$ is dense in $C_0(\mathcal{X})$. This means, for any $\epsilon > 0$ and for any $g \in C_0(\mathcal{X})$, there exists $f \in \mathcal{B}$ such that $\|f - g\|_\infty \le \frac{\epsilon}{2}$. Since $\mathcal{X}$ is an LCH space, $C_0(\mathcal{X})$ is dense in $L^p(\mathcal{X},\mu)$ for all Borel probability measures μ on $\mathcal{X}$ and all $p \in [1,\infty)$. This implies, for any $\epsilon > 0$ and for any $h \in L^p(\mathcal{X},\mu)$, there exists $g \in C_0(\mathcal{X})$ such that $\|g-h\|_{L^p(\mathcal{X},\mu)} \le \frac{\epsilon}{2}$. Consider $\|f-h\|_{L^p(\mathcal{X},\mu)} \le \|f-g\|_{L^p(\mathcal{X},\mu)} + \|g-h\|_{L^p(\mathcal{X},\mu)} \le \frac{\epsilon}{2}+\frac{\epsilon}{2} = \epsilon$, which holds for any ϵ and any $f \in L^p(\mathcal{X},\mu)$. Therefore, $\mathcal{B}$ is dense in $L^p(\mathcal{X},\mu)$ for all Borel probability measures μ on $\mathcal{X}$ and all $p \in [1,\infty)$.

$(\Rightarrow)$ Suppose $\mathcal{B}$ is not dense in $C_0(\mathcal{X})$. Then, by the Hahn-Banach theorem (Theorem 4.7), there exists a $T \in (C_0(\mathcal{X}))'$, $T \neq 0$ such that $T(f) = 0$ for all $f \in \mathcal{B}$. Theorem 7 in [13] shows that for any $T \in (C_0(\mathcal{X}))'$, there exists a probability measure μ on $\mathcal{X}$ and a unique function $h \in L^\infty(\mathcal{X},\mu)$ such that $T(f) = \int_X f(x)h(x)\,d\mu(x)$, $f \in C_0(\mathcal{X})$ with $\|T\| = \|h\|_{L^\infty(\mathcal{X},\mu)}$. Since $T \neq 0$, we have $h \neq 0$. In addition, since μ is a probability measure, $h \in L^q(\mathcal{X},\mu)$, which means there exists $h \neq 0$, $h \in (L^p(\mathcal{X},\mu))'$ such that $\int_X f(x)h(x)\,d\mu(x) = 0$. Therefore, $\mathcal{B}$ is not dense in $L^p(\mathcal{X},\mu)$ for some Borel probability measure μ and any $p \in [1,\infty)$. □

Since it is not easy to check for the denseness of $\mathcal{B}$ in $C_0(\mathcal{X})$ or $L^p(\mathcal{X},\mu)$, in Theorem 7.7, we present an easily checkable characterization for the injectivity of (7.6) when K is bounded continuous and translation invariant on $\mathbb{R}^d$. Note that Theorem 7.7 generalizes Theorem 3.13, which characterizes the injectivity of RKHS embedding (in (3.1)).

Theorem 7.7. *Let $\mathcal{X} = \mathbb{R}^d$. Suppose $K(x,y) = \psi(x-y)$, where ψ is a real-valued function of the form*

$$\psi(x) = \int_{\mathbb{R}^d} e^{i\langle x,\omega\rangle}\,d\Lambda(\omega)$$

and Λ is a finite complex-valued Borel measure on $\mathbb{R}^d$. Then (7.6) is injective if $\operatorname{supp}(\Lambda) = \mathbb{R}^d$. *In addition if K is symmetric, then the converse holds.*

To prove Theorem 7.7, we need many supplementary results which are proven below.

Lemma 7.8. *Let μ be a finite Borel measure and f be a bounded function on $\mathbb{R}^d$. Suppose f is written as*

$$f(x) = \int_{\mathbb{R}^d} e^{i\langle x,\omega\rangle}\,d\Lambda(\omega),$$

with a finite Borel measure Λ on $\mathbb{R}^d$. Then

$$\widehat{f * \mu} = (2\pi)^{d/2}\left(\widehat{\mu}\Lambda\right),$$

where the right hand side is a finite Borel measure[22] *and the equality holds as a tempered distribution.*

Proof. Since the Fourier and inverse Fourier transform give one-to-one correspondence of $\mathcal{S}'_d$, it suffices to show

$$(2\pi)^{-d/2}(f * \mu) = (\widehat{\mu}\Lambda)^{\vee}. \tag{7.8}$$

For an arbitrary $\varphi \in \mathcal{S}_d$,

$$(\widehat{\mu}\Lambda)^{\vee}(\varphi) = (\widehat{\mu}\Lambda)(\varphi^{\vee}) = \int_{\mathbb{R}^d} \varphi^{\vee}(x)\widehat{\mu}(x)\, d\Lambda(x). \tag{7.9}$$

Substituting for $\widehat{\mu}$ and $\varphi^{\vee}$ in Eq. (7.9) and applying Fubini's theorem (Theorem C.1), we have

$$\begin{aligned}(\widehat{\mu}\Lambda)^{\vee}(\varphi) &= (2\pi)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} e^{i\langle \omega - y, x\rangle}\, d\Lambda(x)\right]\varphi(\omega)\, d\omega\, d\mu(y),\\ &= (2\pi)^{-d/2}\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} f(\omega - y)\, d\mu(y)\right]\varphi(\omega)\, d\omega\\ &= (2\pi)^{-d/2}(f * \mu)(\varphi)\end{aligned}$$

and therefore proves (7.8). □

Using Lemma 7.8, in the following, we obtain an alternate representation for $\gamma_K(\mathbb{P}, \mathbb{Q})$—see (7.5)—when K satisfies the assumptions in Theorem 7.7.

Lemma 7.9 (Fourier representation of γ_K)**.** *Suppose K satisfies the conditions in Theorem 7.7. Then*

$$\gamma_K(\mathbb{P}, \mathbb{Q}) = (2\pi)^{d/2}\left\|\left((\overline{\phi}_{\mathbb{P}} - \overline{\phi}_{\mathbb{Q}})\Lambda\right)^{\vee}\right\|_{\mathcal{B}'}, \tag{7.10}$$

where $(\overline{\phi}_{\mathbb{P}} - \overline{\phi}_{\mathbb{Q}})\Lambda$ represents a finite Borel measure defined by (7.7).

[22] Let μ be a finite Borel measure and f be a bounded measurable function on $\mathbb{R}^d$. We then define a finite Borel measure $f\mu$ by

$$(f\mu)(E) = \int_{\mathbb{R}^d} \mathbb{1}_E(x) f(x)\, d\mu(x), \tag{7.7}$$

where E is an arbitrary Borel set and $\mathbb{1}_E$ is its indicator function.

Proof. Consider

$$\int_{\mathbb{R}^d} K(\cdot, x)\, d\mathbb{P}(x) = \int_{\mathbb{R}^d} \psi(\cdot - x)\, d\mathbb{P}(x) = \psi * \mathbb{P}.$$

By Lemma 7.8, we have $\widehat{\psi * \mathbb{P}} = (2\pi)^{d/2}(\widehat{\mathbb{P}}\Lambda)$, which means $\psi * \mathbb{P} = (2\pi)^{d/2}(\widehat{\mathbb{P}}\Lambda)^\vee$, where $\widehat{\mathbb{P}}(\omega) = \int_{\mathbb{R}^d} e^{-i\langle\omega,x\rangle}\, d\mathbb{P}(x)$, $\omega \in \mathbb{R}^d$ (by C.5). Note that $\widehat{\mathbb{P}} = \overline{\phi}_{\mathbb{P}}$. Therefore, substituting for $\int_{\mathbb{R}^d} K(\cdot, x)\, d\mathbb{P}(x)$ in $\gamma_K(\mathbb{P}, \mathbb{Q})$ yields (7.10). □

Lemma 7.10. *Let θ be a bounded continuous function on $\mathbb{R}^d$. Suppose $\theta\Lambda = 0$, where Λ is defined as in Theorem 7.7 and $\theta\Lambda$ is a finite Borel measure defined by (7.7). Then* $\mathrm{supp}(\theta) \subset \mathrm{cl}(\mathbb{R}^d \backslash \mathrm{supp}(\Lambda))$.

Proof. Define $\Omega := \mathrm{supp}(\Lambda)$. Let $W := \{x \in \mathbb{R}^d \,|\, \theta(x) \neq 0\}$. It suffices to show that $W \subset \mathrm{cl}(\mathbb{R}^d \backslash \Omega)$. Suppose W is not contained in $\mathrm{cl}(\mathbb{R}^d \backslash \Omega)$. Then there is a non-empty open subset U such that $U \subset W \cap (\Omega \cup \partial\Omega)$, where $\partial\Omega := \mathrm{cl}(\Omega) \backslash \mathrm{int}(\Omega)$. Fix further a non-empty open subset V with $\mathrm{cl}(V) \subset U$. Since $V \subset \Omega$, there is $\varphi \in \mathcal{D}_d(V)$ with $\Lambda(\varphi) \neq 0$. Take $h \in \mathcal{D}_d(U)$ such that $h = 1$ on $\mathrm{cl}(V)$, and define a continuous function $\varsigma = \frac{h\varphi}{\theta}$ on $\mathbb{R}^d$, which is well-defined from $\mathrm{supp}(h) \subset U$ and $\theta \neq 0$ on U. Since $\theta\Lambda = 0$, by Eq. (7.7), we have

$$\int_{\mathbb{R}^d} \varsigma(x)\theta(x)\, d\Lambda(x) = 0. \tag{7.11}$$

The left hand side of Eq. (7.11) simplifies to

$$\int_{\mathbb{R}^d} \varsigma(x)\theta(x)\, d\Lambda(x) = \int_U \frac{h(x)\varphi(x)}{\theta(x)}\theta(x)\, d\Lambda(x) = \int_U \varphi(x)\, d\Lambda(x) = \Lambda(\varphi) \neq 0,$$

resulting in a contradiction. So, $\mathrm{supp}(\theta) \subset \mathrm{cl}(\mathbb{R}^d \backslash \Omega)$. □

Proof of Theorem 7.7. ($\Leftarrow$) We show that if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, then $\gamma_K(\mathbb{P}, \mathbb{Q})$ is a metric on $M_+^1(\mathcal{X})$, i.e., (7.6) is injective. Let $\gamma_K(\mathbb{P}, \mathbb{Q}) = 0$, which by Lemma 7.9 implies $\big((\overline{\phi}_{\mathbb{P}} - \overline{\phi}_{\mathbb{Q}})\Lambda\big)^\vee = 0$, i.e., $(\overline{\phi}_{\mathbb{P}} - \overline{\phi}_{\mathbb{Q}})\Lambda = 0$. Define $\theta := \overline{\phi}_{\mathbb{P}} - \overline{\phi}_{\mathbb{Q}}$ so that $\theta\Lambda = 0$. By Lemma 7.10, this implies $\mathrm{supp}(\theta) \subset \mathbb{R}^d \backslash \mathrm{supp}(\Lambda)$. Therefore, if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, then $\theta = 0$ a.e., i.e., $\phi_{\mathbb{P}} = \phi_{\mathbb{Q}}$ a.e. Recalling from Theorem C.7 that $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ are uniformly continuous on $\mathbb{R}^d$, we have $\mathbb{P} = \mathbb{Q}$, i.e., γ_K is a metric on $M_+^1(\mathcal{X})$.

$(\Rightarrow)$ Note that since K is real and symmetric, we have that Λ is also real and symmetric, i.e., $\Lambda(d\omega) = \Lambda(-d\omega)$—see Theorem C.7(i). Suppose $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Since the construction of θ in the proof of Theorem 3.13 exploits only the symmetric property of K and does not require K to be pd, the same construction of θ holds which satisfies $\theta, \theta^\vee \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, $\theta(0) = 0$ and $\text{supp}(\theta) = \text{cl}(\mathbb{R}^d \backslash \text{supp}(\Lambda))$. This means, by choosing $\mathbb{Q}$ (with q as its Radon-Nikodym derivative) as in the proof of Theorem 3.13, one can construct a probability measure, $\mathbb{P} \neq \mathbb{Q}$ (with Radon-Nikodym derivative p) as $p := q + \theta^\vee$. However,

$$\gamma_K(\mathbb{P},\mathbb{Q}) = (2\pi)^{d/2} \left\| \left((\overline{\phi}_\mathbb{P} - \overline{\phi}_\mathbb{Q})\Lambda \right)^\vee \right\|_{\mathcal{B}'} = (2\pi)^{d/2} \left\| (\theta\Lambda)^\vee \right\|_{\mathcal{B}'} = 0.$$

Therefore (7.6) is not injective. □

Remark 7.11. *Theorem 7.7 generalizes the characterization for the injectivity of (7.6) when $\mathcal{B}$ is an RKHS. If ψ is a real-valued pd function, then by Bochner's theorem (Theorem 2.1), Λ has to be real, nonnegative and symmetric, i.e., $\Lambda(d\omega) = \Lambda(-d\omega)$. Since ψ need not have to be a pd function for K to be a real, symmetric r.k. of $\mathcal{B}$, Λ need not be nonnegative. More generally, if ψ is a real-valued function on $\mathbb{R}^d$, then Λ is conjugate symmetric, i.e., $\overline{\Lambda(d\omega)} = \Lambda(-d\omega)$. Examples of ψ that are not pd on $\mathbb{R}$ but satisfying the conditions of Theorem 7.7 are: (i) $\psi(x) = \mathbb{1}_{[-\sigma,\sigma]}$, $\sigma > 0$, (ii) $\psi(x) = \exp(-x^2/2\sigma)\sin(\alpha x)$, $\sigma > 0$, $\alpha \neq 0$, etc., for which $\text{supp}(\Lambda) = \mathbb{R}^d$. However, it is not clear whether these are reproducing kernels of some RKBS. In Example 7.18, we provide a construction to generate r.k. that are translation invariant, real and symmetric on $\mathbb{R}^d$ but may not be pd.*

7.2.2 Consistency Analysis

As discussed in Section 5.1, for $\gamma_K(\mathbb{P},\mathbb{Q})$ to be useful in inference applications like two-sample tests, it is required that $\gamma_K(\mathbb{P},\mathbb{Q})$ can be estimated consistently from finite samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$ and the estimator exhibits a *fast* rate of convergence. In Section 5.2, we considered the empirical estimation of $\gamma_K(\mathbb{P},\mathbb{Q})$ when $\mathcal{B}$ is an RKHS (also see [37]) and showed that the empirical estimator, $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ consistently estimates $\gamma_K(\mathbb{P},\mathbb{Q})$ at a convergence rate of $O(m^{-1/2} + n^{-1/2})$, where m (*resp.* n) samples are drawn i.i.d. from $\mathbb{P}$ (*resp.*

$\mathbb{Q}$). The following result (Theorem 7.13) generalizes this consistency result by showing that $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ consistently estimates $\gamma_K(\mathbb{P},\mathbb{Q})$ with a convergence of $O(m^{(1-t)/t}+n^{(1-t)/t})$ if $\mathcal{B}'$ is of *type* t, $1<t\le 2$. Before we present the result, we define the *type* of a Banach space, $\mathcal{B}$ [7, p. 303].

Definition 7.12 (Rademacher type of $\mathcal{B}$)**.** *Let $1\le t\le 2$. A Banach space $\mathcal{B}$ is said to be of t-Rademacher (or, more shortly, of* type t*) if there exists a constant C^* such that for any $N\ge 1$ and any $\{f_j\}_{j=1}^N\subset\mathcal{B}$:*

$$\left(\mathbb{E}\left\|\sum_{j=1}^N \varrho_j f_j\right\|_{\mathcal{B}}^t\right)^{1/t}\le C^*\left(\sum_{j=1}^N \|f_j\|_{\mathcal{B}}^t\right)^{1/t}, \tag{7.12}$$

where $\{\varrho_j\}_{j=1}^N$ are i.i.d. Rademacher (symmetric ±1-valued) random variables.

Clearly, every Banach space is of type 1. Since having type t' for $t'>t$ implies having type t, let us define $t^*(\mathcal{B}):=\sup\{t:\mathcal{B}\text{ has type }t\}$.

Theorem 7.13. *Let $\mathcal{B}$ be an s.i.p. RKBS. Assume $\nu:=\sup_{x\in\mathcal{X}}\sqrt{G(x,x)}<\infty$. Fix $\delta\in(0,1)$. Then with probability $1-\delta$ over the choice of samples $\{X_j^{(1)}\}_{j=1}^m\overset{i.i.d.}{\sim}\mathbb{P}$ and $\{X_j^{(2)}\}_{j=1}^n\overset{i.i.d.}{\sim}\mathbb{Q}$, we have*

$$|\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_K(\mathbb{P},\mathbb{Q})|\le 2C^*\nu\left(m^{\frac{1-t}{t}}+n^{\frac{1-t}{t}}\right)+\sqrt{18\nu^2\log\frac{4}{\delta}}\left(m^{-\frac{1}{2}}+n^{-\frac{1}{2}}\right),$$

where $t=t^*(\mathcal{B}')$ and C^* is some universal constant.

Proof. Since $\gamma_K(\mathbb{P},\mathbb{Q})=\gamma_{\mathcal{F}_K}(\mathbb{P},\mathbb{Q})$ (from Theorem 7.5), the result follows from Theorem 5.11 if we show that

$$R_m(\mathcal{F}_K;\{X_j^{(1)}\}_{j=1}^m)\le C^*\nu m^{(1-t)/t}$$

and

$$R_n(\mathcal{F}_K;\{X_j^{(2)}\}_{j=1}^n)\le C^*\nu n^{(1-t)/t},$$

where $\mathcal{F}_K:=\{f:\|f\|_{\mathcal{B}}\le 1\}$. In the following, we prove this claim. Consider

$$\begin{aligned}R_m(\mathcal{F}_K;\{X_j^{(1)}\}_{j=1}^m) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}_K}\left|\frac{1}{m}\sum_{j=1}^m\varrho_j f(X_j^{(1)})\right|\,\Big|\,\{X_j^{(1)}\}_{j=1}^m\right]\\ &\overset{(7.3)}{=}\mathbb{E}\left[\sup_{f\in\mathcal{F}_K}\left|\left[\frac{1}{m}\sum_{j=1}^m\varrho_j K(\cdot,X_j^{(1)}),f^*\right]_{\mathcal{B}'}\right|\,\Big|\,\{X_j^{(1)}\}_{j=1}^m\right]\end{aligned}$$

$$\begin{aligned}
&\overset{(a_6)}{=} \mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\varrho_j K(\cdot, X_j^{(1)})\right\|_{\mathcal{B}'} \,\Bigg|\, \{X_j^{(1)}\}_{j=1}^{m}\right] \\
&= \mathbb{E}\left[\left(\left\|\frac{1}{m}\sum_{j=1}^{m}\varrho_j K(\cdot, X_j^{(1)})\right\|_{\mathcal{B}'}^{t}\right)^{\frac{1}{t}} \,\Bigg|\, \{X_j^{(1)}\}_{j=1}^{m}\right] \\
&\overset{(*)}{\le} \left(\mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\varrho_j K(\cdot, X_j^{(1)})\right\|_{\mathcal{B}'}^{t} \,\Bigg|\, \{X_j^{(1)}\}_{j=1}^{m}\right]\right)^{\frac{1}{t}} \\
&\overset{(7.12)}{\le} \frac{C^*}{m}\left(\sum_{j=1}^{m}\left\|K(\cdot, X_j^{(1)})\right\|_{\mathcal{B}'}^{t}\right)^{\frac{1}{t}} \\
&\overset{(a_9)}{=} \frac{C^*}{m}\left(\sum_{j=1}^{m}\left\|(G(X_j^{(1)}, \cdot))^*\right\|_{\mathcal{B}'}^{t}\right)^{\frac{1}{t}} \\
&\overset{(a_7)}{=} \frac{C^*}{m}\left(\sum_{j=1}^{m}\left\|G(X_j^{(1)}, \cdot)\right\|_{\mathcal{B}}^{t}\right)^{\frac{1}{t}} \\
&= \frac{C^*}{m}\left(\sum_{j=1}^{m}(G(X_j^{(1)}, X_j^{(1)}))^{\frac{t}{2}}\right)^{\frac{1}{t}} \le C^*\nu m^{\frac{1-t}{t}},
\end{aligned}$$

where we have invoked Jensen's inequality [26, p. 109] in $(*)$. Repeating the similar analysis for $R_n(\mathcal{F}_K; \{X_j^{(2)}\}_{j=1}^{n})$ proves the claim. □

It is clear from Theorem 7.13 that if $t^*(\mathcal{B}') \in (1, 2]$, then $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$ is a consistent estimator of $\gamma_K(\mathbb{P}, \mathbb{Q})$. In addition, the best rate is obtained if $t^*(\mathcal{B}') = 2$, which is the case if $\mathcal{B}$ is an RKHS. In Section 7.3, we will provide examples of s.i.p. RKBSs that satisfy $t^*(\mathcal{B}') = 2$.

7.2.3 Computation of $\gamma_K(\mathbb{P}, \mathbb{Q})$

In (3.5) and (5.13), we showed that $\gamma_K(\mathbb{P}, \mathbb{Q})$ has a nice expression in terms of $K(x, y)$, where $\mathcal{B}$ is an RKHS. We now consider the problem of computing $\gamma_K(\mathbb{P}, \mathbb{Q})$ and $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$ when $\mathcal{B}$ is an s.i.p. RKBS. Consider

$$\gamma_K^2(\mathbb{P}, \mathbb{Q}) = \left\|\int_{\mathcal{X}} K(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot, x)\, d\mathbb{Q}(x)\right\|_{\mathcal{B}'}^{2}$$

$$\begin{aligned}
&\overset{(a_5)}{=} \left[\int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{Q}(x), \int_{\mathcal{X}} K(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\right]_{\mathcal{B}'} \\
&\overset{(a_3)}{=} \left[\int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{P}(x), \int_{\mathcal{X}} K(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\right]_{\mathcal{B}'} \\
&\qquad\qquad - \left[\int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{Q}(x), \int_{\mathcal{X}} K(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\right]_{\mathcal{B}'} \\
&\overset{(7.3)}{=} \int_{\mathcal{X}} \left[K(\cdot,x), \int_{\mathcal{X}} K(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\right]_{\mathcal{B}'} \,d\mathbb{P}(x) \\
&\qquad\qquad - \int_{\mathcal{X}} \left[K(\cdot,x), \int_{\mathcal{X}} K(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\right]_{\mathcal{B}'} \,d\mathbb{Q}(x) \\
&= \int_{\mathcal{X}} \left[K(\cdot,x), \int_{\mathcal{X}} K(\cdot,y)\,d(\mathbb{P}-\mathbb{Q})(y)\right]_{\mathcal{B}'} \,d(\mathbb{P}-\mathbb{Q})(x). \qquad (7.13)
\end{aligned}$$

(7.13) is not reducible as the s.i.p. is not linear in the second argument unless $\mathcal{B}$ is a Hilbert space. This means $\gamma_K(\mathbb{P},\mathbb{Q})$ is not representable in terms of the kernel function, $K(x,y)$ unlike in the case of $\mathcal{B}$ being an RKHS, in which case the s.i.p. in (7.13) reduces to an inner product providing $\gamma_K^2(\mathbb{P},\mathbb{Q}) = \iint_{\mathcal{X}} K(x,y)\,d(\mathbb{P}-\mathbb{Q})(x)\,d(\mathbb{P}-\mathbb{Q})(y)$. Since this issue holds for any $\mathbb{P},\mathbb{Q} \in M_+^1(\mathcal{X})$, it also holds for $\mathbb{P}_m$ and $\mathbb{Q}_n$, which means $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ cannot be computed in a closed form in terms of the kernel, $K(x,y)$ unlike in the case of an RKHS where $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ can be written as a simple V-statistic that depends only on $K(x,y)$ computed at $\{X_j^{(1)}\}_{j=1}^m$ and $\{X_j^{(2)}\}_{j=1}^n$. This is one of the main drawbacks of the RKBS approach where the s.i.p. structure does not allow closed form representations in terms of the kernel K (also see [95] where regularization algorithms derived in RKBS are not easily solvable unlike in an RKHS), and therefore could limit its practical viability. However, in the following section, we present examples of s.i.p. RKBSs for which $\gamma_K(\mathbb{P},\mathbb{Q})$ and $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ can be obtained in closed forms.

7.3 Concrete Examples of RKBS Embeddings

In this section, we present examples of RKBSs and then derive the corresponding $\gamma_K(\mathbb{P},\mathbb{Q})$. To do that, we recall the following result by Zhang et al. [95, Theorem 10].

Theorem 7.14 ([95])**.** *Let $\mathcal{W}$ be an s.i.p. space and $\Phi : \mathcal{X} \to \mathcal{W}$ such that*

$$\mathrm{cl}(\mathrm{span}\,\Phi(\mathcal{X})) = \mathcal{W},\ \mathrm{cl}(\mathrm{span}\,\Phi^*(\mathcal{X})) = \mathcal{W}',$$

where $\Phi^* : \mathcal{X} \to \mathcal{W}'$ is defined as $\Phi^*(x) = (\Phi(x))^*$, $x \in \mathcal{X}$. Then $\mathcal{B} := \{[u, \Phi(\cdot)]_{\mathcal{W}} : u \in \mathcal{W}\}$ equipped with

$$[[u, \Phi(\cdot)]_{\mathcal{W}}\,, [v, \Phi(\cdot)]_{\mathcal{W}}]_{\mathcal{B}} := [u, v]_{\mathcal{W}}$$

and $\mathcal{B}' := \{[\Phi(\cdot), u]_{\mathcal{W}} \,:\, u \in \mathcal{W}\}$ with

$$[[\Phi(\cdot), u]_{\mathcal{W}}\,, [\Phi(\cdot), v]_{\mathcal{W}}]_{\mathcal{B}'} := [v, u]_{\mathcal{W}}$$

are s.i.p. RKBSs, where $\mathcal{B}'$ is the dual of $\mathcal{B}$ with the bilinear form

$$([[u, \Phi(\cdot)]_{\mathcal{W}}\,, [\Phi(\cdot), v]_{\mathcal{W}}])_{\mathcal{B}} := [u, v]_{\mathcal{W}},\ u, v \in \mathcal{W}.$$

Moreover, the s.i.p. kernel G of $\mathcal{B}$ is given by

$$G(x, y) = [\Phi(x), \Phi(y)]_{\mathcal{W}}\,,\ x, y \in \mathcal{X},$$

which coincides with its reproducing kernel, K.

As a corollary to Theorem 7.14, we obtain the following result, which is then used to obtain concrete examples of RKBS embedding.

Corollary 7.15. *Let $(\mathcal{X}, \mathscr{A}, \mu)$ be a measure space. Then for any $1 < p < \infty$, $1 < q < \infty$, $p^{-1} + q^{-1} = 1$,*

$$\mathcal{B}_p(\mathcal{X}) := \left\{ f_u(x) = \int_{\mathcal{X}} u(t) b(x, t)\, d\mu(t) \,:\, u \in L^p(\mathcal{X}, \mu),\ x \in \mathcal{X} \right\}$$

equipped with $[f_u, f_v]_{\mathcal{B}_p} := [u, v]_{L^p(\mathcal{X},\mu)}$, and

$$\mathcal{B}_p'(\mathcal{X}) := \left\{ f_u^*(x) = \int_{\mathcal{X}} \frac{\overline{b(x,t)}|b(x,t)|^{q-2}\overline{u(t)}|u(t)|^{p-2}}{\|b(x,\cdot)\|_{L^q(\mathcal{X},\mu)}^{q-2}\|u\|_{L^p(\mathcal{X},\mu)}^{p-2}}\, d\mu(t) \,:\, \begin{array}{c} u \in L^p(\mathcal{X}, \mu) \\ x \in \mathcal{X} \end{array} \right\}$$

with $[f_u^*, f_v^*]_{\mathcal{B}_p'} := [v, u]_{L^p(\mathcal{X},\mu)}$ are s.i.p. RKBSs with

$$K(x, y) = G(x, y) = \int_{\mathcal{X}} \frac{\overline{b(x,t)}|b(x,t)|^{q-2}}{\|b(x,\cdot)\|_{L^q(\mathcal{X},\mu)}^{q-2}} b(y, t)\, d\mu(t)$$

as the reproducing kernel (also the s.i.p. kernel), where $b(x,\cdot) \in L^q(\mathcal{X},\mu)$, $\forall\, x \in \mathcal{X}$, $\text{cl}(\text{span}\{b(x,\cdot) \,:\, x \in \mathcal{X}\}) = L^q(\mathcal{X},\mu)$ *and* $\text{cl}\left(\text{span}\left\{\frac{\overline{b(x,\cdot)}|b(x,\cdot)|^{q-2}}{\|b(x,\cdot)\|^{q-2}_{L^q(\mathcal{X},\mu)}} : x \in \mathcal{X}\right\}\right) = L^p(\mathcal{X},\mu)$*. Moreover,*

$$\gamma_K(\mathbb{P},\mathbb{Q}) = \left\|\int_{\mathcal{X}} b(x,\cdot)\,d\mathbb{P}(x) - \int_{\mathcal{X}} b(x,\cdot)\,d\mathbb{Q}(x)\right\|_{L^q(\mathcal{X},\mu)}.$$

Proof. Choosing $\mathcal{W} = L^p(\mathcal{X},\mu)$, $\Phi(x) = \frac{\overline{b(x,\cdot)}|b(x,\cdot)|^{q-2}}{\|b(x,\cdot)\|^{q-2}_{L^q(\mathcal{X},\mu)}}$, $\Phi^*(x) = b(x,\cdot)$ in Theorem 7.14 and using $[u,v]_{L^p(\mathcal{X},\mu)}$ as defined in (7.2) proves the result. Now, consider

$$\begin{aligned}
\gamma_K(\mathbb{P},\mathbb{Q}) &= \left\|\int_{\mathcal{X}} K(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\right\|_{\mathcal{B}'} \\
&= \left\|\int_{\mathcal{X}}\int_{\mathcal{X}} \frac{\overline{b(\cdot,t)}|b(\cdot,t)|^{q-2}}{\|b(\cdot,\cdot)\|^{q-2}_{L^q(\mathcal{X},\mu)}} b(x,t)\,d\mu(t)\,d(\mathbb{P}-\mathbb{Q})(x)\right\|_{\mathcal{B}'} \\
&\overset{(*)}{=} \left\|\int_{\mathcal{X}} \frac{\overline{b(\cdot,t)}|b(\cdot,t)|^{q-2}}{\|b(\cdot,\cdot)\|^{q-2}_{L^q(\mathcal{X},\mu)}} \overbrace{\int_{\mathcal{X}} b(x,t)\,d(\mathbb{P}-\mathbb{Q})(x)}^{A(t)}\,d\mu(t)\right\|_{\mathcal{B}'} \\
&= \left\|\frac{\overline{A}|A|^{q-2}}{\|A\|^{q-2}_{L^q(\mathcal{X},\mu)}}\right\|_{L^p(\mathcal{X},\mu)} \\
&= \|A\|_{L^q(\mathcal{X},\mu)},
\end{aligned}$$

where we have invoked Fubini's theorem (Theorem C.1) in $(*)$. □

Corollary 7.15 shows that the embedding of $\mathbb{P}$ into $\mathcal{B}'$ as $\int_{\mathcal{X}} K(\cdot,x)\,d\mathbb{P}(x)$ can be interpreted as embedding $\mathbb{P}$ into $L^q(\mathcal{X},\mu)$ as $\int_{\mathcal{X}} b(x,\cdot)\,d\mathbb{P}(x)$ since these embeddings are isometric. Based on Corollary 7.15, we now present three examples of RKBSs and the corresponding γ_K: Example 7.16 deals with RKBSs induced by a pd kernel and generalizes the distance metric obtained for an RKHS while Examples 7.17 and 7.18 involve RKBS induced by an r.k. that is not pd.

Example 7.16. *Let* μ *be a finite nonnegative Borel measure on* $\mathbb{R}^d$ *and* $b(x,t) = e^{i\langle x,t\rangle}$, $x,t \in \mathbb{R}^d$*, which satisfies the conditions in Corollary 7.15. Therefore, we have*

$$\mathcal{B}_p(\mathbb{R}^d) := \left\{f_u(x) = \int_{\mathbb{R}^d} u(t)e^{i\langle x,t\rangle}\,d\mu(t) \,:\, u \in L^p(\mathbb{R}^d,\mu),\, x \in \mathbb{R}^d\right\}, \tag{7.14}$$

is an RKBS with

$$K(x,y) = G(x,y) = (\mu(\mathbb{R}^d))^{\frac{p-2}{p}} \int_{\mathbb{R}^d} e^{-i\langle x-y,t\rangle}\, d\mu(t) \tag{7.15}$$

as the r.k. and

$$\gamma_K(\mathbb{P},\mathbb{Q}) = \left\| \int_{\mathbb{R}^d} e^{i\langle x,\cdot\rangle}\, d(\mathbb{P}-\mathbb{Q})(x) \right\|_{L^q(\mathbb{R}^d,\mu)} = \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^q(\mathbb{R}^d,\mu)}\,. \tag{7.16}$$

First note that K is a translation invariant pd kernel on $\mathbb{R}^d$ as it is the Fourier transform of a nonnegative finite Borel measure, μ, which follows from Bochner's theorem (Theorem 2.1). Therefore, though the s.i.p. kernel and the r.k. of an RKBS need not have to be symmetric, the space in (7.14) is an interesting example of an RKBS, which is induced by a pd kernel. In particular, it can be seen that many RKBSs ($\mathcal{B}_p(\mathbb{R}^d)$ for any $1 < p < \infty$) have the same r.k (ignoring the scaling factor which can be made one for any p by choosing μ to be a probability measure). Second, note that $\mathcal{B}_p$ is an RKHS when $p = q = 2$ and therefore (7.16) generalizes γ_k (see (3.7)). By Theorem 7.7, it is clear that γ_K in (7.16) is a metric on $M^1_+(\mathbb{R}^d)$ if and only if $\mathrm{supp}(\mu) = \mathbb{R}^d$.

$\mathcal{B}_p(\mathbb{R}^d)$ can also be interpreted as follows. Define

$$\psi(x) = (\mu(\mathbb{R}^d))^{\frac{p-2}{p}} \int_{\mathbb{R}^d} e^{-i\langle x,t\rangle}\, d\mu(t)$$

so that $K(x,y) = \psi(x-y)$. Suppose $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ is strictly pd so that $d\mu(t) = (2\pi)^{-d/2}\widehat{\psi}(t)\, dt$, where $\widehat{\psi}(x) \geq 0,\ \forall\, x \in \mathbb{R}^d$ and $\widehat{\psi} \in L^1(\mathbb{R}^d)$, which follows from Corollary 6.12 in [91]. Then (7.14) can be written as

$$\mathcal{B}_p(\mathbb{R}^d) := \left\{ f_u(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{i\langle x,t\rangle} u(t)\widehat{\psi}(t)\, dt \,:\, u \in L^p(\mathbb{R}^d,\widehat{\psi}),\ x \in \mathbb{R}^d \right\}.$$

Since $\widehat{\psi} \in L^1(\mathbb{R}^d)$ and $u \in L^p(\mathbb{R}^d,\widehat{\psi})$, it is easy to check that $u\widehat{\psi} \in L^1(\mathbb{R}^d)$. Therefore, any $f_u \in \mathcal{B}_p(\mathbb{R}^d)$ can be written as $f_u = (u\widehat{\psi})^{\vee}$, which means $\widehat{f_u} = u\widehat{\psi}$, i.e.,

$$\frac{\widehat{f_u}}{\widehat{\psi}} \in L^p(\mathbb{R}^d,\widehat{\psi}) \Leftrightarrow \frac{\widehat{f_u}}{\widehat{\psi}^{1/q}} \in L^p(\mathbb{R}^d).$$

Therefore (7.14) is equivalent to

$$\mathcal{B}_p(\mathbb{R}^d) := \left\{ f \in C(\mathbb{R}^d) \,:\, \frac{\widehat{f}}{\widehat{\psi}^{1/q}} \in L^p(\mathbb{R}^d) \right\}.$$

By defining $\|f\|_{\mathcal{B}_p} := (2\pi)^{-d/2p}\left\|\frac{\widehat{f}}{\widehat{\psi}^{1/q}}\right\|_{L^p(\mathbb{R}^d)}$ *and using* $\|\cdot\|_{\mathcal{B}_p}$ *in*

$$[f,h]_{\mathcal{B}_p} = \|h\|_{\mathcal{B}_p}\left(\lim_{t\in\mathbb{R},t\to 0}\frac{\|h+tf\|_{\mathcal{B}_p}-\|h\|_{\mathcal{B}_p}}{t} + i\lim_{t\in\mathbb{R},t\to 0}\frac{\|ih+tf\|_{\mathcal{B}_p}-\|h\|_{\mathcal{B}_p}}{t}\right) \tag{7.17}$$

yields

$$[f,g]_{\mathcal{B}_p} = \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\frac{\widehat{f}(\omega)\overline{\widehat{g}(\omega)}|\widehat{g}(\omega)|^{p-2}(\widehat{\psi}(\omega))^{1-p}}{\|g\|^{p-2}_{\mathcal{B}_p}}\,d\omega, \tag{7.18}$$

where we have quoted (7.17) from Proposition 28 of [95]. Note that when $p=q=2$, $\mathcal{B}_p(\mathbb{R}^d)$ *reduces to an RKHS with (7.18) being an inner product (see (2.9)),*

$$\langle f,g\rangle_{\mathcal{B}_2} = (2\pi)^{-d/2}\int_{\mathbb{R}^d}\frac{\widehat{f}(\omega)\overline{\widehat{g}(\omega)}}{\widehat{\psi}(\omega)}\,d\omega.$$

Suppose

$$\psi(x) = \frac{2^{1-s}}{\Gamma(s)}\|x\|_2^{s-d/2}\widetilde{K}_{d/2-s}(\|x\|_2),$$

where $\widetilde{K}$ *represents the modified Bessel function and* $s > d/2$. *Then* $\widehat{\psi}(\omega) = (1+\|\omega\|_2^2)^{-s}$, *which means*

$$\mathcal{B}_p(\mathbb{R}^d) = \left\{f\in C(\mathbb{R}^d) : (1+\|\cdot\|_2^2)^{\frac{s}{q}}\widehat{f}\in L^p(\mathbb{R}^d)\right\} \tag{7.19}$$

represents a Sobolev space of order s. *Note that when* $p=q=2$, *(7.19) reduces to (2.10).*

Example 7.17. *Let* $b(x,t) = e^{\langle x,t\rangle}$, $x,t\in\mathbb{R}^d$ *and* μ *be a finite nonnegative Borel measure such that its moment-generating function, i.e.,* $M^*_\mu(x) := \int_{\mathbb{R}^d} e^{\langle x,t\rangle}\,d\mu(t)$ *exists. Then by Corollary 7.15,*

$$\mathcal{B}_p(\mathbb{R}^d) := \left\{f_u(x) = \int_{\mathbb{R}^d} u(t)e^{\langle x,t\rangle}\,d\mu(t)\;:\; u\in L^p(\mathbb{R}^d,\mu),\, x\in\mathbb{R}^d\right\}$$

is an RKBS with

$$K(x,y) = G(x,y) = \left(M^*_\mu(qx)\right)^{\frac{p-2}{p}} M^*_\mu(x(q-1)+y) \tag{7.20}$$

as the r.k. Suppose $\mathbb{P}$ *and* $\mathbb{Q}$ *are such that* $M^*_\mathbb{P}$ *and* $M^*_\mathbb{Q}$ *exist. Then*

$$\gamma_K(\mathbb{P},\mathbb{Q}) = \left\|\int_{\mathbb{R}^d} e^{\langle x,\cdot\rangle}\,d(\mathbb{P}-\mathbb{Q})(x)\right\|_{L^q(\mathbb{R}^d,\mu)} = \left\|M^*_\mathbb{P} - M^*_\mathbb{Q}\right\|_{L^q(\mathbb{R}^d,\mu)},$$

which is the weighted L^q distance between the moment-generating functions of $\mathbb{P}$ and $\mathbb{Q}$. It is easy to see that if $\operatorname{supp}(\mu) = \mathbb{R}^d$*, then* $\gamma_K(\mathbb{P},\mathbb{Q}) = 0 \Rightarrow M^*_{\mathbb{P}} = M^*_{\mathbb{Q}}$ *a.e.* $\Rightarrow \mathbb{P} = \mathbb{Q}$*, which means* γ_K *is a metric on* $M^1_+(\mathbb{R}^d)$.

Note that K is not symmetric (for $q \neq 2$) and need not be pd. When $p = q = 2$, $K(x,y) = M^*_\mu(x+y)$ is pd and $\mathcal{B}_p(\mathbb{R}^d)$ is an RKHS.

Example 7.18. *Let $b(x,t) = \phi(x-t)$, $x,t \in \mathbb{R}^d$ be real-valued and μ be the Lebesgue measure, where $\phi \in \mathcal{S}_d$ and* $\operatorname{supp}(\widehat{\phi}) = \mathbb{R}^d$*. Then by Corollary 7.15, for any $1 < p < \infty$,*

$$\mathcal{B}_p(\mathbb{R}^d) := \left\{ f_u = \phi * u \,:\, u \in L^p(\mathbb{R}^d) \right\}$$

is an RKBS with

$$\begin{aligned} K(x,y) = G(x,y) &= \|\phi(x-\cdot)\|^{2-q}_{L^q(\mathbb{R}^d)} \int_{\mathbb{R}^d} \phi(x-t)\left|\phi^{q-2}(x-t)\right|\phi(y-t)\,dt \\ &= \|\phi(x-\cdot)\|^{2-q}_{L^q(\mathbb{R}^d)} \int_{\mathbb{R}^d} \phi(t)\,|\phi|^{q-2}(t)\phi(y-x+t)\,dt \qquad (7.21) \\ &\overset{(\star)}{=} (2\pi)^{d/2}\,\|\phi(x-\cdot)\|^{2-q}_{L^q(\mathbb{R}^d)} \left(\widehat{\phi\,|\phi|^{q-2}}\,\phi^\vee\right)^\vee (x-y), \qquad (7.22) \end{aligned}$$

as the r.k. and

$$\gamma_K(\mathbb{P},\mathbb{Q}) = \|\phi * (\mathbb{P}-\mathbb{Q})\|_{L^q(\mathbb{R}^d)}\,,$$

where $(\star)$ follows from Lemma 7.8 and $(\phi * \mathbb{P})(x) := \int_{\mathbb{R}^d} \phi(x-t)\,d\mathbb{P}(t)$ is the convolution of ϕ with $\mathbb{P}$.

Suppose $\phi(x) = \phi(-x)$, $x \in \mathbb{R}^d$ and $\phi(x) \geq 0\,\forall\, x \in \mathbb{R}^d$. Then (7.21) reduces to

$$K(x,y) = G(x,y) = \|\phi(x-\cdot)\|^{2-q}_{L^q(\mathbb{R}^d)}\,(\phi^{q-1} * \phi)(x-y),$$

where $(\phi * u)(x) := \int_{\mathbb{R}^d} \phi(x-t)u(t)\,dt$. Note that if $\|\phi(x-\cdot)\|_{L^q(\mathbb{R}^d)}$ is independent of x, i.e., a constant, then $K(x,y)$ in (7.22) is translation invariant on $\mathbb{R}^d$. Since $\phi(x) = \phi(-x)$, $x \in \mathbb{R}^d$, we have $\eta(x) = \eta(-x)$, $x \in \mathbb{R}^d$, where $\eta := \phi^{q-1} * \phi$ and $\eta \in \mathcal{S}_d$ as $\phi \in \mathcal{S}_d$. This implies $\widehat{\eta} = (2\pi)^{d/2}\widehat{\phi^{q-1}}\widehat{\phi} \in \mathcal{S}_d$ and $\widehat{\eta}(x) = \widehat{\eta}(-x)$, $x \in \mathbb{R}^d$. However, $\widehat{\eta}$ need not be nonnegative everywhere on $\mathbb{R}^d$—if that were the case, then K is pd by Bochner's theorem, as it is the Fourier transform of a nonnegative finite Borel measure, Λ with $d\Lambda(x) = \widehat{\eta}(x)\,dx$—and therefore, K need not be pd though it is symmetric. In the following, we consider two choices for ϕ and obtain K that are pd and symmetric but not pd:

(i) $\phi(x) = (4\pi)^{-d/2} e^{-\frac{\|x\|_2^2}{4}}$, $x \in \mathbb{R}^d$ — *K is pd for any $q \in \mathbb{N}\backslash\{1\}$,*

(ii) $\phi(x) = x^2 e^{-\frac{3x^2}{2}}$, $x \in \mathbb{R}$ — *K is symmetric but not pd.*

It is clear that when $q = 2$, K is a pd kernel as from (7.22), we have $K(x,y) = (2\pi)^{d/2}\left(\widehat{\phi}\phi^\vee\right)^\vee(x-y) = (2\pi)^{d/2}\left(|\widehat{\phi}|^2\right)^\vee(x-y)$ and therefore $\mathcal{B}_2$ is an RKHS.

(i)* $\boldsymbol{\phi(x) = (4\pi)^{-d/2} e^{-\frac{\|x\|_2^2}{4}}}$*: *With this choice of ϕ, we obtain*

$$\mathcal{B}_p(\mathbb{R}^d) := \left\{ f_u(x) = \frac{1}{(4\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{\|x-t\|_2^2}{4}} u(t)\,dt \;:\; u \in L^p(\mathbb{R}^d),\, x \in \mathbb{R}^d \right\}$$

which means any $f_u \in \mathcal{B}_p(\mathbb{R}^d)$ is the Weierstrass transform of u. The r.k. of $\mathcal{B}_p$ is given as

$$K(x,y) = G(x,y) = (4\pi)^{-\frac{d}{p}} q^{-\frac{d}{q}} e^{-\frac{\|x-y\|_2^2}{4p}}, \tag{7.23}$$

which is clearly pd for any $q \in \mathbb{N}\backslash\{1\}$.

(ii)* $\boldsymbol{\phi(x) = x^2 e^{-\frac{3x^2}{2}}}$*: *With this choice of ϕ and $q = 3$, we obtain*

$$\mathcal{B}_{\frac{3}{2}}(\mathbb{R}) := \left\{ f_u(x) = \int_{\mathbb{R}} (x-t)^2 e^{-\frac{3(x-t)^2}{2}} u(t)\,dt \;:\; u \in L^{\frac{3}{2}}(\mathbb{R}),\, x \in \mathbb{R} \right\}$$

as an RKBS with r.k.

$$K(x,y) = G(x,y) = \frac{e^{-(x-y)^2}}{243}\left(\frac{4\pi^2}{25}\right)^{\frac{1}{6}} \left(4(x-y)^6 + 9(x-y)^4 - 18(x-y)^2 + 15\right) \tag{7.24}$$

which is not pd (though it is symmetric on $\mathbb{R}$) as its Fourier transform given by

$$\left(\widehat{\phi}^2\,\widehat{\phi}\right)(x) = \frac{-e^{-\frac{(x-y)^2}{4}}}{34992\sqrt{2}} \left(x^6 - 39x^4 + 216x^2 - 324\right) \tag{7.25}$$

is not nonnegative at every $x \in \mathbb{R}$.

Refer to Appendix B for the derivation of (7.23)–(7.25). Consider

$$\gamma_K(\mathbb{P},\mathbb{Q}) = \left\| \int_{\mathbb{R}^d} \phi(\cdot - x)\,d(\mathbb{P}-\mathbb{Q})(x) \right\|_{L^q(\mathbb{R}^d)} \overset{(\star)}{=} \left\| \left(\widehat{\phi}\,(\overline{\phi_{\mathbb{P}}} - \overline{\phi_{\mathbb{Q}}})\right)^\vee \right\|_{L^q(\mathbb{R}^d)},$$

where we have derived the equality in ($\star$) by using the idea in Lemma 7.9. Since $\operatorname{supp}(\widehat{\phi}) = \mathbb{R}^d$*, we have $\gamma_K(\mathbb{P},\mathbb{Q}) = 0 \Rightarrow (\widehat{\phi}\,(\overline{\phi_{\mathbb{P}}} - \overline{\phi_{\mathbb{Q}}}))^\vee = 0 \Rightarrow \widehat{\phi}\,(\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}) = 0 \Rightarrow \phi_{\mathbb{P}} = \phi_{\mathbb{Q}}$ a.e., which implies $\mathbb{P} = \mathbb{Q}$ and therefore γ_K is a metric on $M_+^1(\mathbb{R}^d)$.*

Having presented concrete examples of RKBSs and the corresponding Banach space embedding of probability measures, we now consider the problem of computing $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ for K derived in Examples 7.16-7.18. In Section 7.2.3, we showed that $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ does not have a nice closed form expression unlike in the case of $\mathcal{B}$ being an RKHS. However, in the following, we show that if K is the kernel in Examples 7.16 or 7.17, then for certain choices of q, $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ has a closed form expression.

For the kernel, K in (7.15), let us consider the estimation of $\gamma_K(\mathbb{P},\mathbb{Q})$:

$$\begin{aligned}\gamma_K(\mathbb{P}_m,\mathbb{Q}_n) &= \|\phi_{\mathbb{P}_m}-\phi_{\mathbb{Q}_m}\|_{L^q(\mathbb{R}^d,\mu)}\\ &= \left(\int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} e^{i\langle x,t\rangle}\,d(\mathbb{P}_m-\mathbb{Q}_n)(x)\right|^q d\mu(t)\right)^{\frac{1}{q}}\\ &= \left(\int_{\mathbb{R}^d}\left|\frac{1}{m}\sum_{j=1}^m e^{i\langle t,X_j^{(1)}\rangle}-\frac{1}{n}\sum_{j=1}^n e^{i\langle t,X_j^{(2)}\rangle}\right|^q d\mu(t)\right)^{\frac{1}{q}}. \qquad (7.26)\end{aligned}$$

For $\mathcal{B}_p(\mathbb{R}^d)$ is defined in (7.14), since $\mathcal{B}'_p(\mathbb{R}^d)$ is of type $\min(q,2)$ for $1\le q\le\infty$ [7, p. 304], by Theorem 7.13, $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$ estimates $\gamma_K(\mathbb{P},\mathbb{Q})$ consistently at a convergence rate of $O(m^{\frac{\max(1-q,-1)}{\min(q,2)}}+n^{\frac{\max(1-q,-1)}{\min(q,2)}})$ for $q\in(1,\infty)$, with the best rate of $O(m^{-1/2}+n^{-1/2})$ attainable when $q\in[2,\infty)$. Now, the problem reduces to computing $\gamma_K(\mathbb{P}_m,\mathbb{Q}_n)$. Since computing (7.26) in closed form may not be possible for all q, (7.26) can be approximated as

$$\gamma_K^a(\mathbb{P}_m,\mathbb{Q}_n)=(\mu(\mathbb{R}^d))^{1/q}\left(\frac{1}{N}\sum_{s=1}^N\left|\frac{1}{m}\sum_{j=1}^m e^{i\langle t_s,X_j^{(1)}\rangle}-\frac{1}{n}\sum_{j=1}^n e^{i\langle t_s,X_j^{(2)}\rangle}\right|^q\right)^{\frac{1}{q}}, \qquad (7.27)$$

where $\{t_s\}_{s=1}^N$ are N random samples drawn i.i.d. from the probability measure, $\eta:=\mu/\mu(\mathbb{R}^d)$. However, when $q=2$, (7.26) can be computed very efficiently in closed form (in terms of K) as a V-statistic [37], given by

$$\gamma_K^2(\mathbb{P}_m,\mathbb{Q}_n)=\sum_{j,l=1}^m\frac{K(X_j^{(1)},X_l^{(1)})}{m^2}+\sum_{j,l=1}^n\frac{K(X_j^{(2)},X_l^{(2)})}{n^2}-2\sum_{j=1}^m\sum_{l=1}^n\frac{K(X_j^{(1)},X_l^{(2)})}{mn},$$

rather than through (7.27). More generally, we show below that if $q=2s$, $s\in\mathbb{N}$, then (7.26) can be written in terms of the kernel, $K(x,y)$ as

$$\gamma_K^q(\mathbb{P}_m,\mathbb{Q}_n)=\int_{\mathbb{R}^d}|\phi_{\mathbb{P}_m}(t)-\phi_{\mathbb{Q}_n}(t)|^q\,d\mu(t)$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^d} (\phi_{\mathbb{P}_m} - \phi_{\mathbb{Q}_n})(t)\overline{(\phi_{\mathbb{P}_m} - \phi_{\mathbb{Q}_n})(t)} \overset{s}{\cdots} \\
&\qquad\qquad (\phi_{\mathbb{P}_m} - \phi_{\mathbb{Q}_n})(t)\overline{(\phi_{\mathbb{P}_m} - \phi_{\mathbb{Q}_n})(t)}\, d\mu(t) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i\langle t,x_1\rangle}\, d(\mathbb{P}_m - \mathbb{Q}_n)(x_1) \int_{\mathbb{R}^d} e^{-i\langle t,x_2\rangle}\, d(\mathbb{P}_m - \mathbb{Q}_n)(x_2) \overset{s}{\cdots} \\
&\qquad\qquad \int_{\mathbb{R}^d} e^{i\langle t,x_{q-1}\rangle}\, d(\mathbb{P}_m - \mathbb{Q}_n)(x_{q-1}) \int_{\mathbb{R}^d} e^{-i\langle t,x_q\rangle}\, d(\mathbb{P}_m - \mathbb{Q}_n)(x_q)\, d\mu(t) \\
&\overset{(\star)}{=} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i\langle t,x_1-x_2\rangle}\, d(\mathbb{P}_m - \mathbb{Q}_n)(x_1)\, d(\mathbb{P}_m - \mathbb{Q}_n)(x_2)\right) \overset{s}{\cdots} \\
&\qquad\qquad \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i\langle t,x_{q-1}-x_q\rangle}\, d(\mathbb{P}_m - \mathbb{Q}_n)(x_{q-1})\, d(\mathbb{P}_m - \mathbb{Q}_n)(x_q)\right) d\mu(t) \\
&\overset{(\star)}{=} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \overset{q}{\cdots} \int_{\mathbb{R}^d} e^{i\langle t,\sum_{j=1}^s x_{2j-1} - \sum_{j=1}^s x_{2j}\rangle} \prod_{j=1}^q d(\mathbb{P}_m - \mathbb{Q}_n)(x_j)\right) d\mu(t) \\
&\overset{(\star)}{=} \int_{\mathbb{R}^d} \overset{q}{\cdots} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} e^{i\langle t,\sum_{j=1}^q (-1)^{j-1}x_j\rangle}\, d\mu(t)\right) \prod_{j=1}^q d(\mathbb{P}_m - \mathbb{Q}_n)(x_j) \\
&\overset{(\star\star)}{=} \frac{(\mu(\mathbb{R}^d))^{\frac{2}{p}}}{(\mu(\mathbb{R}^d))} \int_{\mathbb{R}^d} \overset{q}{\cdots} \int_{\mathbb{R}^d} K\left(\sum_{j=1}^s x_{2j-1}, \sum_{j=1}^s x_{2j}\right) \prod_{j=1}^q d(\mathbb{P}_m - \mathbb{Q}_n)(x_j), \qquad (7.28)
\end{aligned}
$$

where we have invoked Fubini's theorem in $(\star)$ and (7.15) in $(\star\star)$. Note that choosing $s = 1$ results in (5.13). (7.28) shows that $\gamma_K^q(\mathbb{P}_m, \mathbb{Q}_n)$ which can be computed in a closed form in terms of K at a complexity of $O(2^q m^q)$, assuming $m = n$, which means the least complexity is obtained for $q = 2$. Therefore, for appropriate choices of q, the RKBS embedding in Example 7.16 is useful in practice as $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$ is consistent and has a closed form expression. However, the drawback of the RKBS framework is that the computation of $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$ is more involved than its RKHS counterpart.

Based on the discussion so far on the computation of $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$ for K in (7.15), similar ideas can be used to address the problem of computing $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$ for the kernel, K in (7.20). Akin to (7.26), we get

$$
\begin{aligned}
\gamma_K(\mathbb{P}_m, \mathbb{Q}_n) &= \left\| M^*_{\mathbb{P}_m} - M^*_{\mathbb{Q}_m} \right\|_{L^q(\mathbb{R}^d,\mu)} \\
&= \left(\int_{\mathbb{R}^d} \left|\int_{\mathbb{R}^d} e^{\langle x,t\rangle}\, d(\mathbb{P}_m - \mathbb{Q}_n)(x)\right|^q d\mu(t)\right)^{\frac{1}{q}} \\
&= \left(\int_{\mathbb{R}^d} \left|\frac{1}{m}\sum_{j=1}^m e^{\langle t,X_j^{(1)}\rangle} - \frac{1}{n}\sum_{j=1}^n e^{\langle t,X_j^{(2)}\rangle}\right|^q d\mu(t)\right)^{\frac{1}{q}},
\end{aligned}
$$

which cannot be computed in a closed form and therefore one can resort to an approximation similar to the one in (7.27). Instead, if $q = 2s$, $s \in \mathbb{N}$, it can be shown that $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$ can be written in a closed form similar to the one in (7.28) but in terms of M^*_μ as

$$\gamma_K(\mathbb{P}_m, \mathbb{Q}_n) = \left(\int_{\mathbb{R}^d} \overset{q}{\cdots} \int_{\mathbb{R}^d} M^*_\mu \left(\sum_{j=1}^{q} x_j \right) \prod_{j=1}^{q} d(\mathbb{P}_m - \mathbb{Q}_n)(x_j) \right)^{\frac{1}{q}},$$

which requires $O(2^q m^q)$ computations, assuming $m = n$ (the least complexity is obtained at $q = 2$). Therefore, the kernel and the corresponding RKBS in Example 7.17 share the similar advantages and disadvantages of their counterparts in Example 7.16.

In the above, while we derived closed form expressions for $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$, where K is chosen from Examples 7.16 and 7.17 and $q = 2s$, $s \in \mathbb{N}$, it is can be shown that if K is chosen from Example 7.18, such a closed expression is not possible for $\gamma_K(\mathbb{P}_m, \mathbb{Q}_n)$ unless $q = 2$.

7.4 Discussion

In this chapter, we have generalized the notion of RKHS embedding of probability measures to Banach spaces, in particular by embedding probability measures into RKBS that are uniformly Fréchet differentiable and uniformly convex. Similar to the case of RKHS, these embeddings are shown to be determined by the reproducing kernel (r.k.) of the RKBS and then characterized its injectivity. We also showed that the Rademacher type of the RKBS determines the consistency and rate of convergence of the empirical estimator of the distance between the embeddings of probability measures $\mathbb{P}$ and $\mathbb{Q}$. Although one of the drawbacks of the RKBS approach (compared to its RKHS counterpart) is the non-computability of this estimator in a closed form, in general, we presented non-trivial examples of RKBS for which the closed form computation is possible.

Many issues are still open in the theory of RKBS and its associated distribution embeddings. First, unlike in the case of RKHS where the r.k. is positive definite, there is no such nice characterization for the r.k. of an RKBS. Therefore, a systematic study is needed to characterize the properties of RKBS and its r.k. Second, though we provided a few examples of RKBS, one of which being a variation of the Sobolev space, it is interesting to consider generalizations of other Banach spaces like Besov, Orlicz, Orlicz-Sobolev spaces, etc., and study the corresponding embeddings.

Bibliographic Notes

This chapter is based on joint unpublished work with Kenji Fukumizu and Gert Lanckriet. The dissertation author was the primary investigator.

A Relation Between IPMs and ϕ-Divergences

In this appendix, we discuss the relation between IPMs and ϕ-divergences and show that IPMs are essentially different from ϕ-divergences.

Based on the definitions of IPM and ϕ-divergence, it is clear that $\{\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) : \mathcal{F}\}$ and $\{D_\phi(\mathbb{P},\mathbb{Q}) : \phi\}$ represent classes of IPMs and ϕ-divergences (on $\mathbb{P}$ and $\mathbb{Q}$) indexed by $\mathcal{F}$ and ϕ, respectively. Let us define $\mathscr{P}_\lambda(\mathcal{X})$ as the set of all probability measures, $\mathbb{P}$ that are absolutely continuous with respect to some σ-finite measure, λ on $\mathcal{X}$. For $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda(\mathcal{X})$, let $p = \frac{d\mathbb{P}}{d\lambda}$ and $q = \frac{d\mathbb{Q}}{d\lambda}$ be the Radon-Nikodym derivatives of $\mathbb{P}$ and $\mathbb{Q}$ with respect to λ. For $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda$ (so that $\mathbb{P} \ll \mathbb{Q}$), it is easy to check that the above two classes intersect at $\mathcal{F} = \{f : \|f\|_\infty \le 1\}$ and $\phi(t) = |t-1|$, i.e., $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q}) = \int_{\mathcal{X}} |p-q|\, d\lambda$, which is the total-variation distance. So, a natural question to consider is for what conditions on $\mathcal{F}$ and ϕ is $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ for all $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda(\mathcal{X})$? This shows the degree of overlap between the class of IPMs and the class of ϕ-divergences. We answer this in the following theorem, where we show that the total-variation distance is the only "non-trivial"[23] IPM that is also a ϕ-divergence.

Theorem A.1 (Necessary and sufficient conditions)**.** *Suppose $\mathcal{F}_\star$ be the set of all real-valued measurable functions on $\mathcal{X}$ and Φ be the class of all convex functions $\phi : [0,\infty) \to (-\infty,\infty]$ continuous at 0 and finite on $(0,\infty)$. Let $\mathcal{F} \subset \mathcal{F}_\star$ and $\phi \in \Phi$. Then for any $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda(\mathcal{X})$, $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ if and only if any one*

[23] Choosing $\mathcal{F}$ to be the set of all real-valued measurable functions on $\mathcal{X}$ and $\phi(t) = 0$ if $t = 1$ and $+\infty$ if $t \neq 1$ yields $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q}) = 0$ if $\mathbb{P} = \mathbb{Q}$ and $+\infty$ if $\mathbb{P} \neq \mathbb{Q}$. It is easy so show that the converse also holds. This choice of $\mathcal{F}$ and ϕ shows that IPM is trivially a ϕ-divergence.

of the following hold:

(i) $\mathcal{F} = \{f : \|f\|_\infty \le \frac{\beta-\alpha}{2}\}, \phi(u) = \alpha(u-1)\mathbb{1}_{[0,1]}(u) + \beta(u-1)\mathbb{1}_{[1,\infty)}(u)$ *for some* $\alpha < \beta < \infty$, *i.e.,* $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q}) = \frac{\beta-\alpha}{2}\int_{\mathcal{X}} |p-q|\,d\lambda$.

(ii) $\mathcal{F} = \{f : f = c, c \in \mathbb{R}\}, \phi(u) = \alpha(u-1)\mathbb{1}_{[0,\infty)}(u)$, $\alpha \in \mathbb{R}$, *i.e.,* $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q}) = 0$.

Proof. Define $\mathbb{P}f := \int_{\mathcal{X}} f\,d\mathbb{P}$.

($\Leftarrow$) Suppose (i) holds. Then for any $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda(\mathcal{X})$, we have

$$\begin{aligned}\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) &= \sup\left\{|\mathbb{P}f - \mathbb{Q}f| : \|f\|_\infty \le \frac{\beta-\alpha}{2}\right\}\\ &= \frac{\beta-\alpha}{2}\sup\{|\mathbb{P}f - \mathbb{Q}f| : \|f\|_\infty \le 1\}\\ &= \frac{\beta-\alpha}{2}\int_{\mathcal{X}} |p-q|\,d\lambda \overset{(a)}{=} D_\phi(\mathbb{P},\mathbb{Q}),\end{aligned}$$

where (a) follows from simple algebra after substituting ϕ in $D_\phi(\mathbb{P},\mathbb{Q})$ (see [43]). This means $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ and $D_\phi(\mathbb{P},\mathbb{Q})$ are equal to the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$.

Suppose (ii) holds. Then $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = 0$ and $D_\phi(\mathbb{P},\mathbb{Q}) = \alpha\int_{\mathcal{X}} q\,\phi(p/q)\,d\lambda = \alpha\int_{\mathcal{X}}(p-q)\,d\lambda = 0$.

($\Rightarrow$) Suppose $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ for any $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda(\mathcal{X})$. Since $\gamma_{\mathcal{F}}$ is a pseudometric on $\mathscr{P}_\lambda(\mathcal{X})$ (irrespective of $\mathcal{F}$), D_ϕ is a pseudometric on $\mathscr{P}_\lambda(\mathcal{X})$. Through the simple modification of Theorem 2 in [43], it can be shown that if D_ϕ is a pseudometric then $\phi(u) = \alpha(u-1)\mathbb{1}_{[0,1]}(u) + \beta(u-1)\mathbb{1}_{[1,\infty)}(u)$ for some $\beta \ge \alpha$, which means for $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda(\mathcal{X})$, $D_\phi(\mathbb{P},\mathbb{Q}) = \frac{\beta-\alpha}{2}\int_{\mathcal{X}}|p-q|\,d\lambda$ if $\beta > \alpha$ and $D_\phi(\mathbb{P},\mathbb{Q}) = 0$ if $\beta = \alpha$. Now, let us consider two cases.

Case 1: $\beta > \alpha$

Since $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ for all $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda(\mathcal{X})$, we have $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \frac{\beta-\alpha}{2}\int_{\mathcal{X}}|p-q|\,d\lambda = \frac{\beta-\alpha}{2}\sup\{|\mathbb{P}f - \mathbb{Q}f| : \|f\|_\infty \le 1\} = \sup\{|\mathbb{P}f - \mathbb{Q}f| : \|f\|_\infty \le \frac{\beta-\alpha}{2}\}$ and therefore $\mathcal{F} = \{f : \|f\|_\infty \le \frac{\beta-\alpha}{2}\}$.

Case 2: $\beta = \alpha$

$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}}|\mathbb{P}f - \mathbb{Q}f| = 0$ for all $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda(\mathcal{X})$, which means $\forall\,\mathbb{P},\mathbb{Q} \in$

$\mathscr{P}_\lambda(\mathcal{X})$, $\forall\, f \in \mathcal{F}$, $\mathbb{P}f = \mathbb{Q}f$. This, in turn, means f is a constant on $\mathcal{X}$, i.e., $\mathcal{F} = \{f : f = c,\, c \in \mathbb{R}\}$. □

Note that in Theorem A.1, the cases (i) and (ii) are disjoint as $\alpha < \beta$ in case (i) and $\alpha = \beta$ in case (ii). Case (i) shows that the family of ϕ-divergences and the family of IPMs intersect only at the total variation distance. Case (ii) is trivial as the distance between any two probability measures is zero. This result shows that IPMs and ϕ-divergences are essentially different.

B Derivation of (7.23)–(7.25)

For ϕ defined in Example 7.18, K is obtained as

$$K(x,y) = (2\pi)^{d/2}\,\|\phi(x-\cdot)\|^{2-q}_{L^q(\mathbb{R}^d)}\left(\widehat{\phi\,|\phi|^{q-2}}\phi^\vee\right)^\vee(x-y), \tag{B.1}$$

which is the (inverse) Fourier transform of $\widehat{\phi\,|\phi|^{q-2}}\phi^\vee$. If $\phi(x)=\phi(-x),\ \forall\, x\in\mathbb{R}^d$ and $\phi(x)\geq 0,\ \forall\, x\in\mathbb{R}^d$, then (B.1) reduces to

$$K(x,y) = (2\pi)^{d/2}\,\|\phi(x-\cdot)\|^{2-q}_{L^q(\mathbb{R}^d)}\left(\widehat{\phi^{q-1}}\widehat{\phi}\right)^\vee(x-y). \tag{B.2}$$

To derive (7.23)–(7.25), we use the following identities, where $\alpha>0$.

$$\begin{aligned}
\int_{\mathbb{R}^d} e^{-\alpha\|x\|_2^2} &= \left(\frac{\pi}{\alpha}\right)^{\frac{d}{2}}\\
\int_{\mathbb{R}} (x-b)^{2r}e^{-\alpha(x-b)^2} &= \sqrt{\frac{\pi}{\alpha}}\frac{1}{(2\alpha)^r}\frac{(2r)!}{r!\,2^r},\ \ r\in\mathbb{N}\\
\widehat{e^{-\alpha\|x\|_2^2}} &= \frac{1}{(2\alpha)^{d/2}}e^{-\frac{\|x\|_2^2}{4\alpha}}\\
\frac{d^2}{dx^2}e^{-\alpha x^2} &= \alpha\left(4\alpha x^2-2\right)e^{-\alpha x^2}\\
\frac{d^4}{dx^4}e^{-\alpha x^2} &= \alpha^2\left(16\alpha^2x^4-48\alpha x^2+12\right)e^{-\alpha x^2}\\
\frac{d^6}{dx^6}e^{-\alpha x^2} &= \alpha^3\left(64\alpha^3x^6-480\alpha^2x^4+720\alpha x^2-120\right)e^{-\alpha x^2}\\
\widehat{x^nf(x)} &= i^n\,\frac{d^n}{dx^n}\widehat{f(x)}
\end{aligned}$$

B.1 Derivation of (7.23)

Since $\phi(x)=(4\pi)^{-d/2}e^{-\|x\|_2^2/4}$, we have

$$\|\phi(x-\cdot)\|^{2-q}_{L^q(\mathbb{R}^d)} = \left(\int_{\mathbb{R}^d}\frac{1}{(4\pi)^{dq/2}}e^{-\frac{q\|x-t\|_2^2}{4}}\,dt\right)^{\frac{2-q}{q}} = \left(\frac{(4\pi)^{d(1-q)/2}}{q^{d/2}}\right)^{\frac{2-q}{q}}, \tag{B.3}$$

$$\left(\widehat{\phi^{q-1}}\widehat{\phi}\right)(x) = (4\pi)^{-dq/2} e^{-\widehat{\frac{(q-1)\|x\|_2^2}{4}}} e^{-\widehat{\frac{\|x\|_2^2}{4}}} = \frac{1}{(4\pi)^{dq/2}} \left(\frac{4}{(q-1)}\right)^{d/2} e^{-\frac{q}{q-1}\|x\|_2^2},$$

and

$$\left(\widehat{\phi^{q-1}}\widehat{\phi}\right)^{\vee}(x) = \frac{1}{(4\pi)^{dq/2}} \left(\frac{4}{(q-1)}\right)^{d/2} e^{\widehat{-\frac{q}{q-1}\|x\|_2^2}} = \frac{1}{(4\pi)^{dq/2}} \left(\frac{2}{q}\right)^{d/2} e^{-\frac{q-1}{4q}\|x\|_2^2}. \tag{B.4}$$

Therefore, using (B.3) and (B.4) in (B.2) yields,

$$K(x,y) = (4\pi)^{\frac{d(1-q)}{q}} q^{-\frac{d}{q}} e^{-\frac{q-1}{4q}\|x-y\|_2^2} = (4\pi)^{-\frac{d}{p}} q^{-\frac{d}{q}} e^{-\frac{\|x-y\|_2^2}{4p}},$$

where we used $p^{-1} + q^{-1} = 1$.

B.2 Derivation of (7.24) and (7.25)

For $\phi(x) = x^2 e^{-\frac{3x^2}{2}}$, we have

$$\|\phi(x-\cdot)\|_{L^q(\mathbb{R}^d)}^{2-q} = \left(\int_{\mathbb{R}} (x-t)^{2q} e^{-\frac{3q(x-t)^2}{2}}\,dt\right)^{\frac{2-q}{q}} = \left(\sqrt{\frac{2\pi}{3q}}\frac{(2q)!}{q!(6q)^q}\right)^{\frac{2-q}{q}}, \tag{B.5}$$

$$\begin{aligned}\widehat{\phi^{q-1}}(x) = \widehat{x^{2(q-1)}e^{-\frac{3(q-1)x^2}{2}}} &= (-1)^{q-1}\frac{d^{2q-2}}{dx^{2q-2}} e^{-\widehat{\frac{3(q-1)x^2}{2}}} \\ &= \frac{(-1)^{q-1}}{\sqrt{3(q-1)}}\frac{d^{2q-2}}{dx^{2q-2}} e^{-\frac{x^2}{6(q-1)}},\end{aligned} \tag{B.6}$$

and

$$\widehat{\phi}(x) = -3^{-1/2}\frac{d^2}{dx^2}e^{-\frac{x^2}{6}} = \frac{3-x^2}{9\sqrt{3}}e^{-\frac{x^2}{6}}.$$

Since (B.6) is not computable in a closed form and q has to be an integer greater than 2 (as $q = 2$ yields a pd kernel, K), let us choose $q = 3$ for the ease of computation. Then, (B.5) and (B.6) reduce to

$$\|\phi(x-\cdot)\|_{L^3(\mathbb{R}^d)}^{-1} = 9\,(50\pi)^{-\frac{1}{6}} \tag{B.7}$$

$$\widehat{\phi^2}(x) = \frac{1}{\sqrt{6}}\frac{d^4}{dx^4}e^{-\frac{x^2}{12}} = \frac{x^4 - 36x^2 + 108}{1296\sqrt{2}}e^{-\frac{x^2}{12}}$$

and therefore

$$\begin{aligned}\left(\widehat{\phi}^2\widehat{\phi}\right)(x) &= \frac{(3-x^2)(x^4-36x^2+108)}{2^{\frac{9}{2}}3^7}e^{-\frac{x^2}{4}} \\ &= \frac{-(x^6-39x^4+216x^2-324)}{2^{\frac{9}{2}}3^7}e^{-\frac{x^2}{4}},\end{aligned} \tag{B.8}$$

$$\begin{aligned}\left(\widehat{\phi}^2\widehat{\phi}\right)^{\vee}(x) &= \frac{-1}{2^{\frac{9}{2}}3^7}(x^6-39x^4\widehat{+216x^2}-324)e^{-\frac{x^2}{4}} \\ &= \frac{1}{2^{\frac{9}{2}}3^7}\left(\frac{d^6}{dx^6}+39\frac{d^4}{dx^4}+216\frac{d^2}{dx^2}+324\right)\widehat{e^{-\frac{x^2}{4}}} \\ &= \frac{1}{2^4 3^7}\left(\frac{d^6}{dx^6}+39\frac{d^4}{dx^4}+216\frac{d^2}{dx^2}+324\right)e^{-x^2} \\ &= \frac{e^{-x^2}}{2^4 3^7}\Big((64x^6-480x^4+720x^2-120)+39(16x^4-48x^2+12) \\ &\qquad\qquad +216(4x^2-2)+324\Big) \\ &= \frac{(4x^6+9x^4-18x^2+15)}{3^7}e^{-x^2}.\end{aligned} \tag{B.9}$$

Using (B.7) and (B.9) in (B.2) yields

$$K(x,y) = \frac{e^{-(x-y)^2}}{243}\left(\frac{4\pi^2}{25}\right)^{\frac{1}{6}}\left(4(x-y)^6+9(x-y)^4-18(x-y)^2+15\right). \tag{B.10}$$

Note that K in (B.10) is real and symmetric but is not pd as its Fourier transform, $\widehat{\phi}^2\widehat{\phi}$ in (B.8) is not nonnegative everywhere on $\mathbb{R}$—for example $\left(\widehat{\phi}^2\widehat{\phi}\right)(6) = -0.4398$.

C Appendix

In this appendix, we summarize several results and notions that are used in the dissertation.

C.1 Definitions

In the following, we define various notions that are used in the dissertation.

C.1.1 Standard Spaces

$\boldsymbol{C(\mathcal{X})}$**,** $\boldsymbol{C_b(\mathcal{X})}$**,** $\boldsymbol{C_0(\mathcal{X})}$**,** $\boldsymbol{\|\cdot\|_\infty}$ **:** Let $\mathcal{X}$ be a topological space. $C(\mathcal{X})$ denotes the space of all *continuous functions* on $\mathcal{X}$. $C_b(\mathcal{X})$ is the space of all *bounded, continuous functions* on $\mathcal{X}$. For a locally compact Hausdorff space (examples include $\mathbb{R}^d$, infinite discrete sets, topological manifolds, etc.), $\mathcal{X}$, $f \in C(\mathcal{X})$ is said to *vanish at infinity* if for every $\epsilon > 0$ the set $\{x : |f(x)| \geq \epsilon\}$ is compact.[24] The class of all continuous f on $\mathcal{X}$ which vanish at infinity is denoted as $C_0(\mathcal{X})$. The spaces $C_b(\mathcal{X})$ and $C_0(\mathcal{X})$ are endowed with the *supremum norm*, $\|\cdot\|_\infty$ defined as $\|f\|_\infty := \sup_{x\in\mathcal{X}} |f(x)|$ for $f \in C_0(\mathcal{X}) \subset C_b(\mathcal{X})$. The space of all r-continuously differentiable functions on $\mathcal{X}$ is denoted by $C^r(\mathcal{X})$, $0 \leq r \leq \infty$.

$\boldsymbol{L^p(\mathcal{X},\mu)}$**,** $\boldsymbol{L^p(\mathcal{X})}$**,** $\boldsymbol{\|\cdot\|_{L^p(\mathcal{X},\mu)}}$ **:** For a measure space, $(\mathcal{X}, \mathscr{A}, \mu)$, $L^p(\mathcal{X},\mu)$ denotes the Banach space of p-power ($p \geq 1$) μ-integrable functions endowed with the L^p*-norm*

$$\|f\|_{L^p(\mathcal{X},\mu)} := \left(\int_{\mathcal{X}} |f|^p \, d\mu\right)^{\frac{1}{p}}.$$

[24] LCH spaces have a rich supply of continuous functions that vanish outside compact sets—see Tietze extension theorem [26, Theorem 4.34].

We use $L^p(\mathcal{Z})$ for $L^p(\mathcal{Z},\mu)$ and dx for $d\mu(x)$ if μ is the Lebesgue measure on $\mathcal{Z} \subset \mathbb{R}^d$.

$\mathcal{D}_d$, $\mathcal{S}_d$: Let $\mathcal{D}_d$ be the space of *compactly supported infinitely differentiable functions* on $\mathbb{R}^d$, i.e.,

$$\mathcal{D}_d = \{f \in C^\infty(\mathbb{R}^d) \,|\, \mathrm{supp}(f) \text{ is bounded}\},$$

where $\mathrm{supp}(f) = \mathrm{cl}\left(\{x \in \mathbb{R}^d \,|\, f(x) \neq 0\}\right)$. A function $f : \mathbb{R}^d \to \mathbb{C}$ is said to *decay rapidly, or be rapidly decreasing*, if for all $N \in \mathbb{N}$,

$$\sup_{\|\alpha\|_1 \leq N} \sup_{x \in \mathbb{R}^d} (1 + \|x\|_2^2)^N |(T_\alpha f)(x)| < \infty,$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is an ordered d-tuple of nonnegative α_j, $\|\alpha\|_1 = \sum_{j=1}^d \alpha_j$ and $T_\alpha = \left(\frac{1}{i}\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{1}{i}\frac{\partial}{\partial x_d}\right)^{\alpha_d}$. $\mathcal{S}_d$, called the *Schwartz class*, denotes the vector space of rapidly decreasing functions. Note that $\mathcal{D}_d \subset \mathcal{S}_d$. Also, for any $p \in [1,\infty]$, $\mathcal{S}_d \subset L^p(\mathbb{R}^d)$.

$M_b(\mathcal{X})$, $M_+^b(\mathcal{X})$, $M_+^1(\mathcal{X})$: A signed Borel measure μ on a topological space $\mathcal{X}$ is said to be *finite* if $\|\mu\| := |\mu|(\mathcal{X}) < \infty$, where $|\mu|$ is the *total-variation* of μ. $M_+^b(\mathcal{X})$ denotes the space of all *finite Borel measures* on $\mathcal{X}$ while $M_b(\mathcal{X})$ denotes the space of all *finite signed Borel measures* on $\mathcal{X}$. The space of all *Borel probability measures* is denoted as $M_+^1(\mathcal{X}) := \{\mu \in M_+^b(\mathcal{X}) : \mu(\mathcal{X}) = 1\}$. For $\mu \in M_b(\mathcal{X})$, the *support* of μ is defined as

$$\mathrm{supp}(\mu) = \{x \in \mathcal{X} \,|\, \text{for any open set } U \text{ such that } x \in U,\, |\mu|(U) \neq 0\}. \qquad \text{(C.1)}$$

$M_{bc}(\mathcal{X})$ denotes the space of all *compactly supported finite signed Borel measures* on $\mathcal{X}$. A signed *Radon measure* μ on a Hausdorff space $\mathcal{X}$ is a Borel measure on $\mathcal{X}$ satisfying

(*i*) $\mu(C) < \infty$ for each compact subset $C \subset \mathcal{X}$,

(*ii*) $\mu(B) = \sup\{\mu(C) \,|\, C \subset B,\, C \,\text{compact}\}$ for each B in the Borel σ-algebra of $\mathcal{X}$.

If $\mathcal{X}$ is a Polish space, then by Ulam's theorem, every finite Borel measure is Radon [23, Theorem 7.1.4].

Lip$(\mathcal{X},\rho)$, BL$(\mathcal{X},\rho)$, $\|\cdot\|_L$, $\|\cdot\|_{BL}$: A real-valued function, f on a metric space, $(\mathcal{X},\rho)$ is said to be *L-Lipschitz* if $|f(x)-f(y)|\le L\rho(x,y)$, $\forall\, x,y\in\mathcal{X}$, where $L>0$. The *Lipschitz semi-norm* of f, denoted as $\|f\|_L$ is defined as

$$\begin{aligned}\|f\|_L &:= \inf\{L : |f(x)-f(y)|\le L\rho(x,y),\,\forall\, x,y\in\mathcal{X}\}\\ &= \sup\left\{\frac{|f(x)-f(y)|}{\rho(x,y)} : x\ne y \text{ in } \mathcal{X}\right\}.\end{aligned}$$

The space of all Lipschitz functions on $\mathcal{X}$ is defined as

$$\text{Lip}(\mathcal{X},\rho) := \{f:\mathcal{X}\to\mathbb{R}\,|\,\|f\|_L<\infty\}.$$

The space of all *bounded Lipschitz functions* is defined as

$$BL(\mathcal{X},\rho) := \{f:\mathcal{X}\to\mathbb{R}\,|\,\|f\|_{BL}<\infty\},$$

where $\|f\|_{BL}:=\|f\|_L+\|f\|_\infty$ is the *dual-bounded Lipschitz norm.*

C.1.2 Distributions and Fourier Transforms

For $f\in L^1(\mathbb{R}^d)$, $\widehat{f}$ and $f^\vee$ represent the *Fourier transform* and *inverse Fourier transform* of f respectively, defined as

$$\widehat{f}(y) := \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{-i\langle y,x\rangle} f(x)\,dx,\ y\in\mathbb{R}^d, \tag{C.2}$$

$$f^\vee(x) := \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{i\langle x,y\rangle} f(y)\,dy,\ x\in\mathbb{R}^d. \tag{C.3}$$

Since $\mathcal{S}_d\in L^p(\mathbb{R}^d)$ for any $p\in[1,\infty]$, the above definitions of $\widehat{f}$ and $f^\vee$ hold for any $f\in\mathcal{S}_d$. It can be shown that for any $f\in\mathcal{S}_d$, $\widehat{f}\in\mathcal{S}_d$ and $f^\vee\in\mathcal{S}_d$ (see [26, Chapter 9] and [65, Chapter 6] for details). For $f\notin L^1(\mathbb{R}^d)$ but $f\in L^2(\mathbb{R}^d)$, the Fourier transform $\widehat{f}$ is defined to be the limit, in the L^2-norm, of the sequence $\{\widehat{f_n}\}$ of Fourier transforms of any sequence $\{f_n\}$ of functions belonging to $\mathcal{S}_d$, such that f_n converges in the L^2-norm to the given function $f\in L^2(\mathbb{R}^d)$, as $n\to\infty$. The function $\widehat{f}$ is defined almost everywhere on $\mathbb{R}^d$ and belongs to $L^2(\mathbb{R}^d)$. See [32, Chapter IV, Lesson 22] for details.

Distributions, $\mathcal{D}'_d$: A linear functional on $\mathcal{D}_d$ which is continuous with respect to the Fréchet topology [65, Definition 6.3] is called a *distribution* in $\mathbb{R}^d$. The space of all distributions in $\mathbb{R}^d$ is denoted by $\mathcal{D}'_d$.

As examples, if f is *locally integrable* on $\mathbb{R}^d$ (this means that f is Lebesgue measurable and $\int_K |f(x)|\,dx < \infty$ for every compact $K \subset \mathbb{R}^d$), then the functional D_f defined by

$$D_f(\varphi) = \int_{\mathbb{R}^d} f(x)\varphi(x)\,dx,\ \varphi \in \mathcal{D}_d, \tag{C.4}$$

is a distribution. Similarly, if μ is a Borel measure on $\mathbb{R}^d$, then

$$D_\mu(\varphi) = \int_{\mathbb{R}^d} \varphi(x)\,d\mu(x),\ \varphi \in \mathcal{D}_d,$$

defines a distribution D_μ in $\mathbb{R}^d$, which is identified with μ.

Support of a distribution: For an open set $U \subset \mathbb{R}^d$, $\mathcal{D}_d(U)$ denotes the subspace of $\mathcal{D}_d$ consisting of the functions with support contained in U. Suppose $D \in \mathcal{D}'_d$. If U is an open set of $\mathbb{R}^d$ and if $D(\varphi) = 0$ for every $\varphi \in \mathcal{D}_d(U)$, then D is said to *vanish* or be *null* in U. Let W be the union of all open $U \subset \mathbb{R}^d$ in which D vanishes. The complement of W is the *support* of D.

Tempered distributions, $\mathcal{S}'_d$ and Fourier transform on $\mathcal{S}'_d$: A linear continuous functional (with respect to the Fréchet topology) over the space $\mathcal{S}_d$ is called a *tempered distribution* and the space of all tempered distributions in $\mathbb{R}^d$ is denoted by $\mathcal{S}'_d$. For example, every compactly supported distribution is tempered.

For any $f \in \mathcal{S}'_d$, the Fourier and inverse Fourier transforms are defined as

$$\widehat{f}(\varphi) := f(\widehat{\varphi}),\ \varphi \in \mathcal{S}_d,$$
$$f^\vee(\varphi) := f(\varphi^\vee),\ \varphi \in \mathcal{S}_d,$$

respectively. The Fourier transform is a linear, one-to-one, bicontinuous mapping from $\mathcal{S}'_d$ to $\mathcal{S}'_d$.

For complete details on distribution theory and Fourier transforms of distributions, we refer the reader to [26, Chapter 9] and [65, Chapter 6].

C.2 Supplementary Results

For completeness, we present supplementary results that are used to prove the main results in this dissertation.

C.2.1 Real Analysis

Theorem C.1 (Fubini [26, Theorem 2.37]). *If $f \in L^1(\mathcal{X}, \mu \otimes \nu)$, then*

$$\iint_{\mathcal{X}} f(x,y)\, d\mu(x)\, d\nu(y) = \iint_{\mathcal{X}} f(x,y)\, d\nu(y)\, d\mu(x).$$

Theorem C.2 (Riesz Representation Theorem for Hilbert Space [26, Theorem 5.25]). *Suppose $\mathcal{H}$ is an Hilbert space. If $T \in \mathcal{H}'$, there is a unique $g \in \mathcal{H}$ such that $T(f) = \langle f, g\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.*

Theorem C.3 (Riesz Representation Theorem for $C_0(\mathcal{X})$ [26, Theorem 7.17]). *Let $\mathcal{X}$ be an LCH space, and for $\mu \in \mathscr{M}_b(\mathcal{X})$ and $f \in C_0(\mathcal{X})$ let $T_\mu(f) = \int_{\mathcal{X}} f\, d\mu$, where $\mathscr{M}_b(\mathcal{X})$ is the set of all finite signed Radon measures on $\mathcal{X}$. Then the map $\mu \mapsto T_\mu$ is an isometric isomorphism from $\mathscr{M}_b(\mathcal{X})$ to $(C_0(\mathcal{X}))'$.*

Lemma C.4 (Lipschitz extension [52,92]). *Let $(\mathcal{X}, \rho)$ be a metric and $f : A \to \mathbb{R}$, $A \subset \mathcal{X}$, be an L-Lispchitz function. Then there exists an L-Lipschitz function $F : \mathcal{X} \to \mathbb{R}$ such that $F|A = f$. In particular, F can be explicitly constructed as*

$$F(x) = \alpha \min_{a\in A}(f(x_i) + L\rho(x,a)) + (1-\alpha)\max_{a\in A}(f(x_i) - L\rho(x,a)),$$

for any $\alpha \in [0,1]$.

Lemma C.5 (Bounded Lipschitz extension [23, Proposition 11.2.3]). *If $A \subset \mathcal{X}$ and $f \in BL(A,\rho)$, then f can be extended to a function $h \in BL(\mathcal{X},\rho)$ with $h = f$ on A and $\|h\|_{BL} = \|f\|_{BL}$. Additionally, it is possible to explicitly construct h as*

$$h = \max\left(-\|f\|_\infty, \min\left(g, \|f\|_\infty\right)\right),$$

where g is a function on $\mathcal{X}$ such that $g = f$ on A and $\|g\|_L = \|f\|_L$.

The following result characterizes strictly pd kernels on $\mathbb{T}$, which we quote from [53]. Before we state the result, we introduce some notation. For natural numbers m and n and a set A of integers, $m + nA := \{j \in \mathbb{Z} \,|\, j = m + na,\, a \in A\}$. An increasing sequence $\{c_l\}$ of nonnegative integers is said to be *prime* if it is not contained in any set of the form $p_1\mathbb{N} \cup p_2\mathbb{N} \cup \cdots \cup p_n\mathbb{N}$, where $p_1, p_2, \ldots, p_n$ are prime numbers. Any infinite increasing sequence of prime numbers is a trivial example of a prime sequence. We write $\mathbb{N}_n^0 := \{0, 1, \ldots, n\}$.

Theorem C.6 ([53]). *Let ψ be a pd function on $\mathbb{T}$ of the form in (2.4). Let $\overline{N} := \{|n| : A_\psi(n) > 0,\, n \in \mathbb{Z}\} \subset \mathbb{N} \cup \{0\}$. Then ψ is strictly pd if $\overline{N}$ has a subset of the form $\cup_{l=0}^{\infty}(b_l + c_l\mathbb{N}_l^0)$, in which $\{b_l\} \cup \{c_l\} \subset \mathbb{N}$ and $\{c_l\}$ is a prime sequence.*

C.2.2 Fourier Analysis

Theorem C.7 (Fourier transform of a measure). *Let μ be a finite Borel measure on $\mathbb{R}^d$. The Fourier transform of μ is given by*

$$\widehat{\mu}(\omega) = \int_{\mathbb{R}^d} e^{-i\langle \omega, x\rangle}\, d\mu(x),\ \omega \in \mathbb{R}^d, \tag{C.5}$$

which is a bounded, uniformly continuous function on $\mathbb{R}^d$. In addition, $\widehat{\mu}$ satisfies the following properties:

(i) $\overline{\widehat{\mu}(\omega)} = \widehat{\mu}(-\omega)$, $\forall\, \omega \in \mathbb{R}^d$, that is, $\widehat{\mu}$ is conjugate symmetric,

(ii) $\widehat{\mu}(0) = 1$.

(iii) (Bochner's Theorem) $\widehat{\mu}$ is pd if and only if $\mu \in M_b^+(\mathbb{R}^d)$.

Lemma C.8 (Riemann-Lebesgue [65, Theorem 7.5]). *If $f \in L^1(\mathbb{R}^d)$, then $\widehat{f} \in C_0(\mathbb{R}^d)$, and $\|\widehat{f}\|_\infty \le \|f\|_1$.*

Theorem C.9 (Paley-Wiener [65, Theorem 7.23]). *If $f \in \mathcal{D}'_d$ has compact support, then $\widehat{f}$ is the restriction to $\mathbb{R}^d$ of an entire function on $\mathbb{C}^d$.*

C.2.3 Convex Analysis

Theorem C.10 ([64, Theorem 32.1]). *Let f be a convex function, and let C be a convex set contained in the domain of f. If f attains its supremum relative to C at some point of relative interior of C, then f is actually constant throughout C.*

C.2.4 Concentration Inequalities and Empirical Processes

Theorem C.11 (Almost sure convergence of an empirical process [85, Theorem 3.7]). *Let $F(x) = \sup_{f\in\mathcal{F}} |f(x)|$ be the envelope function for $\mathcal{F}$. Assume that*

$\int F\, d\mathbb{P} < \infty$, and that for any $\varepsilon > 0$,

$$\frac{1}{m}\mathcal{H}(\mathcal{F}, L^1(\mathbb{P}_m), \varepsilon) \xrightarrow{\mathbb{P}} 0.$$

Then $\sup_{f\in\mathcal{F}} \int f\, d(\mathbb{P}_m - \mathbb{P}) \xrightarrow{a.s.} 0$.

Theorem C.12 (McDiarmid's Inequality [51])**.** *Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be independent random variables taking values in a set $\mathcal{X}$, and assume that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies*

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_n)| \leq c_j, \tag{C.6}$$

$\forall\, x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathcal{X}$. Then for every $\epsilon > 0$,

$$Pr(f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n) \geq \epsilon) \leq e^{\frac{-2\epsilon^2}{\sum_{j=1}^n c_j^2}}. \tag{C.7}$$

Lemma C.13 (Symmetrization [86])**.** *Let $\varrho_1, \ldots, \varrho_N$ be i.i.d. Rademacher random variables. Suppose $\{X_j\}_{j=1}^N \overset{i.i.d.}{\sim} \mathbb{P}$. Then,*

$$\mathbb{E} \sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_N f| \leq 2\, \mathbb{E} \sup_{f\in\mathcal{F}} \left| \frac{1}{N} \sum_{j=1}^{N} \varrho_j f(X_j) \right|. \tag{C.8}$$

Bibliography

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B (Methodological)*, 28:131–142, 1966.

[2] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, UK, 1999.

[3] M. A. Arcones and E. Giné. Limit theorems for U-processes. *Annals of Probability*, 21:1494–1542, 1993.

[4] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[5] I. Guyon B. E. Boser and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Madison, WI, 1992. ACM.

[6] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[7] B. Beauzamy. *Introduction to Banach spaces and their Geometry*. North-Holland, The Netherlands, 1985.

[8] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Spring Verlag, New York, 1984.

[9] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, London, UK, 2004.

[10] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] P. Brémaud. *Mathematical Principles of Signal Processing*. Springer-Verlag, New York, 2001.

[12] A. Caponnetto, M. Pontil, C. Micchelli, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

[13] C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

[14] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[15] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarium Mathematicarum Hungarica*, 2:299–318, 1967.

[16] F. Cucker and D-X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, UK, 2007.

[17] W. Dahmen and C. A. Micchelli. Some remarks on ridge functions. *Approx. Theory Appl.*, 3:139–143, 1987.

[18] V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer-Verlag, NY, 1999.

[19] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez. Testing of goodness of fit based on the L_2-Wasserstein distance. *Annals of Statistics*, 27:1230–1239, 1999.

[20] R. Der and D. Lee. Large-marign classification in Banach spaces. In *JMLR Workshop and Conference Proceedings*, volume 2, pages 91–98. AISTATS, 2007.

[21] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. Wiley, New York, 1985.

[22] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[23] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, UK, 2002.

[24] N. Dunford and J. T. Schwartz. *Linear operators. I: General theory*. Wiley-Interscience, New York, 1958.

[25] A. A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's inequality. *IEEE Trans. Information Theory*, 49(6):1491–1498, 2003.

[26] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, New York, 1999.

[27] B. Fuglede and F. Topsøe. Jensen-Shannon divergence and Hilbert space embedding, 2003. Preprint.

[28] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[29] K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(5):1871–1905, 2009.

[30] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

[31] K. Fukumizu, B. K. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 473–480, 2009.

[32] C. Gasquet and P. Witomski. *Fourier Analysis and Applications.* Springer-Verlag, New York, 1999.

[33] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[34] J. R. Giles. Classes of semi-inner-product spaces. *Trans. Amer. Math. Soc.*, 129:436–446, 1967.

[35] R. M. Gray, D. L. Neuhoff, and P. C. Shields. A generalization of Ornstein's $\overline{d}$ distance with applications to information theory. *Annals of Probability*, 3:315–328, 1975.

[36] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. Technical Report 157, MPI for Biological Cybernetics, 2007.

[37] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

[38] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, 2008.

[39] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N. Logothetis. Kernel constrained covariance for dependence measurement. In Z. Ghahramani and R. Cowell, editors, *Proc. 10^{th} International Workshop on Artificial Intelligence and Statistics*, pages 1–8, 2005.

[40] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In Z. Ghahramani and R. Cowell, editors, *Proc. 10^{th} International Workshop on Artificial Intelligence and Statistics*, pages 136–143, 2005.

[41] M. Hein, T.N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in SVMs. In *Proceedings of the 26th DAGM Symposium*, pages 270–277, Berlin, 2004. Springer.

[42] E. Hewitt. Linear functionals on spaces of continuous functions. *Fundamenta Mathematicae*, 37:161–189, 1950.

[43] M. Khosravifard, D. Fooladivanda, and T. A. Gulliver. Confliction of the convexity and metric properties in f-divergences. *IEICE Trans. Fundamentals*, E90-A(9):1848–1853, 2007.

[44] G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[45] A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional space. *American Mathematical Society Translations*, 2(17):277–364, 1961.

[46] G. R. G. Lanckriet, N. Christianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:24–72, 2004.

[47] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypothesis.* Springer-Verlag, New York, 2005.

[48] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Trans. Information Theory*, 52(10):4394–4412, 2006.

[49] T. Lindvall. *Lectures on the Coupling Method.* John Wiley & Sons, New York, 1992.

[50] G. Lumer. Semi-inner-product spaces. *Trans. Amer. Math. Soc.*, 100:29–43, 1961.

[51] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.

[52] E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40:837–842, 1934.

[53] V. A. Menegatto. Strictly positive definite kernels on the circle. *Rocky Mountain Journal of Mathematics*, 25(3):1149–1163, 1995.

[54] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

[55] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

[56] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. In *IEEE International Symposium on Information Theory*, 2007.

[57] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. Technical Report 764, Department of Statistics, University of California, Berkeley, 2008.

[58] A. Pinkus. Strictly positive definite functions on a real inner product space. *Adv. Comput. Math.*, 20:263–271, 2004.

[59] S. T. Rachev. On a class of minimum functionals in a space of probability measures. *Theory of Probability and its Applications*, 29:41–48, 1984.

[60] S. T. Rachev. The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory of Probability and its Applications*, 29:647–676, 1985.

[61] S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. John Wiley & Sons, Chichester, 1991.

[62] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Vol. I Theory, Vol. II Applications*. Probability and its Applications. Springer-Verlag, Berlin, 1998.

[63] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[64] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[65] W. Rudin. *Functional Analysis*. McGraw-Hill, USA, 1991.

[66] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman, Harlow, UK, 1988.

[67] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proc. of the 14th Annual Conference on Learning Theory*, pages 416–426, 2001.

[68] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[69] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980.

[70] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, UK, 2004.

[71] G. R. Shorack. *Probability for Statisticians*. Springer-Verlag, New York, 2000.

[72] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proc. 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, Berlin, Germany, 2007.

[73] L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th International Conference on Machine Learning*, pages 992–999, 2008.

[74] N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In G. Lugosi and H. U. Simon, editors, *Proc. of the 19th Annual Conference on Learning Theory*, pages 169–183, 2006.

[75] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1750–1758. MIT Press, 2009.

[76] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. Non-parametric estimation of integral probability metrics. In *Proc. IEEE International Symposium on Information Theory*, pages 1428–1432, June 2010.

[77] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In Y. W. Teh and M. Titterington, editors, *Proc. 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Workshop and Conference Proceedings*. JMLR, 2010.

[78] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. S chölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[79] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In R. Servedio and T. Zhang, editors, *Proc. of the 21^{st} Annual Conference on Learning Theory*, pages 111–122, 2008.

[80] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[81] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, 2008.

[82] J. Stewart. Positive definite functions and generalizations, an historical survey. *Rocky Mountain Journal of Mathematics*, 6(3):409–433, 1976.

[83] I. Vajda. *Theory of Statistical Inference and Information.* Kluwer Academic Publishers, Boston, 1989.

[84] S. S. Vallander. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.*, 18:784–786, 1973.

[85] S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, Cambridge, UK, 2000.

[86] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer-Verlag, New York, 1996.

[87] V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[88] U. von Luxburg and O. Bousquet. Distance-based classification with Lipschitz functions. *Journal for Machine Learning Research*, 5:669–695, 2004.

[89] Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Trans. Information Theory*, 51(9):3064–3074, 2005.

[90] Q. Wang, S. R. Kulkarni, and S. Verdú. A nearest-neighbor approach to estimating divergence between continuous random vectors. In *IEEE Symposium on Information Theory*, 2006.

[91] H. Wendland. *Scattered Data Approximation.* Cambridge University Press, Cambridge, UK, 2005.

[92] H. Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36:63–89, 1934.

[93] Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In *Proc. of the 22nd Annual Conference on Learning Theory*, 2009.

[94] Y. Ying and D. X. Zhou. Learnability of Gaussians with flexible variances. *Journal of Machine Learning Research*, 8:249–276, 2007.

[95] H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

[96] V. M. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.

Notation and Symbols

Sets

$\emptyset$	empty set
$\mathbb{N}$	set of positive integers
$\mathbb{N}_n$	$\{1, 2, \ldots, n\}$
$\mathbb{Z}$	set of positive or negative integers including 0
$\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_{++}$	set of real, nonnegative real and positive real numbers
$\mathbb{T}$	$[0, 2\pi)$
$\mathbb{C}$	set of complex numbers
$\mathrm{int}(A)$, $\mathrm{cl}(A)$	interior and closure of a set A
$A \cup B$, $A \cap B$	union and intersection of A and B
$\|A\|$	number of elements of a set A
$A \times B$	cartesian product of A and B

Functions

$\mathbb{1}_A(x)$	indicator function, $\mathbb{1}_A(x) = 1$, if $x \in A$, 0 otherwise
$\lceil \cdot \rceil$	$\lceil x \rceil = \min\{y \in \mathbb{Z} : y \geq x\}$, $x \in \mathbb{R}$
id	identity map $x \mapsto x$
$\phi_{\mathbb{P}}$, $\phi_{\mathbb{Q}}$	characteristic functions of probability measures $\mathbb{P}$, $\mathbb{Q}$
M^*_μ	moment generating function of measure μ
$f * g$	convolution of f and g: $\int_{\mathbb{R}^d} f(\cdot - x) g(x)\, dx$
$f * \mathbb{P}$	convolution of f and $\mathbb{P}$: $\int_{\mathbb{R}^d} f(\cdot - x)\, d\mathbb{P}(x)$
$\widehat{f}$, $f^\vee$	Fourier and inverse Fourier transforms of f
$a \vee b$	$\max(a, b)$
$\mathrm{sign}(x)$	$\mathbb{1}_{x>0}(x) - \mathbb{1}_{x<0}(x)$

Spaces

$\mathcal{X}$	space of input values
$\mathcal{B}$	RKBS or a generic Banach space
$\mathcal{B}_p(\mathcal{X})$	Reproducing kernel Banach space
$\mathcal{B}'$, $\mathcal{B}'_p(\mathcal{X})$	topological duals of $\mathcal{B}$ and $\mathcal{B}_p(\mathcal{X})$
$\mathcal{F}$, $\mathcal{G}$	arbitrary class of functions
$\mathcal{F}_\star$	set of all measurable functions on $\mathcal{X}$

$\mathcal{F}_W$, $\mathcal{F}_\beta$, $\mathcal{F}_k$	unit balls w.r.t. $\|\cdot\|_L$, $\|\cdot\|_{BL}$ and $\|\cdot\|_{\mathcal{H}}$
$\mathcal{H}$	RKHS or generic Hilbert space
$\mathcal{H}_k$	Reproducing kernel Hilbert space
$C(\mathcal{X})$	space of continuous functions $f : \mathcal{X} \to \mathbb{R}$
$C_b(\mathcal{X})$	space of bounded continuous functions $f : \mathcal{X} \to \mathbb{R}$
$C^r(\mathcal{X})$	space of r-differentiable functions, $0 \leq r \leq \infty$
$C_0(\mathcal{X})$	space of functions $f : \mathcal{X} \to \mathbb{R}$ vanishing at infinity
$L^p(\mathcal{X}, \mu)$	space of p-power $(p \geq 1)$ μ-integrable functions
$L^p(\mathcal{X})$	space of p-power Lebesgue integrable functions on $\mathcal{X} \subset \mathbb{R}^d$
$\ell_p(\mathcal{X})$, ℓ_p	space of p-summable functions or sequences
$\mathcal{D}_d$	space of compactly supported $f \in C^\infty(\mathbb{R}^d)$
$\mathcal{S}_d$	space of rapidly decaying functions on $\mathbb{R}^d$
$\mathcal{D}'_d$, $\mathcal{S}'_d$	Distributions and tempered distributions on $\mathbb{R}^d$
$\mathrm{Lip}(\mathcal{X}, \rho)$	space of Lipschitz functions on a metric space $(\mathcal{X}, \rho)$
$\mathrm{BL}(\mathcal{X}, \rho)$	space of bounded Lipschitz functions on $(\mathcal{X}, \rho)$

Norms and other symbols

$\|\cdot\|$, $\|\cdot\|_2$	Euclidean norm
$\|\cdot\|_p$	p-norm, $\|x\|_p := (\sum_{j=1}^d \lvert x_j\rvert^p)^{1/p}$ for $x \in \mathbb{C}^d$
$\|\cdot\|_{L^p(\mathcal{X},\mu)}$	L^p-norm
$\|\cdot\|_\infty$	supremum norm
$\|\cdot\|_{\mathcal{B}}$, $\|\cdot\|_{\mathcal{B}'}$	norms of RKBSs, $\mathcal{B}$ and $\mathcal{B}'$
$\|\cdot\|_{\mathcal{H}}$, $\|\cdot\|_{\mathcal{H}_k}$	norm of RKHS $\mathcal{H}$
$\|\cdot\|_L$	Lipschitz semi-norm
$\|\cdot\|_{BL}$	dual-bounded Lipschitz norm
$(\cdot,\cdot)_{\mathcal{B}}$	bilinear form on $\mathcal{B} \times \mathcal{B}'$
$\langle\cdot,\cdot\rangle$, $\langle\cdot,\cdot\rangle_{\mathcal{H}}$	inner product (in Hilbert space $\mathcal{H}$)
$[\cdot,\cdot]_{\mathcal{B}}$	semi-inner-product (in s.i.p. space $\mathcal{B}$)

Measures and Random variables

$(\mathcal{X}, \mathscr{A})$	measurable space with σ-algebra $\mathscr{A}$
μ	unspecified measure, sometimes signed measure
$\mu \otimes \nu$	product measure of μ and ν
$\mu \ll \nu$	μ is absolutely continuous w.r.t. ν
$\mu \perp \nu$	μ and ν are singular
$\widehat{\mu}$	Fourier transform of μ
$\lvert\mu\rvert$	total variation of μ
$\operatorname{supp}\mu$	support of μ (also defined for functions)
λ	Lebesgue measure on $\mathbb{R}$ (also on $\mathbb{R}^d$)
δ_x	Dirac measure at $x \in \mathcal{X}$
$\mathbb{P}$, $\mathbb{Q}$	probability measures
$\mathbb{P}_m$, $\mathbb{Q}_n$	empirical estimators of $\mathbb{P}$ and $\mathbb{Q}$

$\mathbb{E}(\cdot)$	expectation operator
$\mathbb{P}f$	$\int_{\mathcal{X}} f \, d\mathbb{P}$
ϱ	Rademacher random variable
X, Y, $X_j^{(1)}$, $X_j^{(2)}$	random variables
$R_m(\mathcal{F}; \{x_j\}_{j=1}^m)$	Rademacher complexity of $\mathcal{F}$
$U_m(\mathcal{F}; \{x_j\}_{j=1}^m)$	Rademacher chaos complexity of $\mathcal{F}$
$\overset{a.s.}{\to}$, $\overset{\mathbb{P}}{\to}$, $\overset{w}{\to}$	convergence in (a.s., $\mathbb{P}$, weak) sense
$O_{\mathbb{P}}(\cdot)$	$X_n = O_{\mathbb{P}}(r_n)$: X_n/r_n is bounded in probability
$M_+^b(\mathcal{X})$	set of all finite Borel measures on $\mathcal{X}$
$M_b(\mathcal{X})$	set of all finite signed Borel measures on $\mathcal{X}$
$M_{bc}(\mathcal{X})$	set of all compactly supported $\mu \in M_b(\mathcal{X})$
$M_+^1(\mathcal{X})$	set of all Borel probability measures on $\mathcal{X}$

Metrics/Divergences on $M_+^1(\mathcal{X})$

$\gamma_{\mathcal{F}}$	integral probability metric
γ_k, γ_K, γ	MMD and generalized MMD
W, W_1	Kantorovich and Wasserstein distances
β	Dudley metric
TV	Total variation distance
D_ϕ	ϕ-divergence

Kernels

k, K	reproducing kernels of RKHS and RKBS
$\mathcal{K}$	family of positive definite kernels
δ_x	Dirac functional
ψ	$k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$
Λ	Fourier transform of ψ
A_ψ	Fourier series coefficients of ψ when $\mathcal{X} = \mathbb{T}^d$

Miscellaneous

$a := b$, $b =: a$	a is defined by b
d	dimension of the vector
δ	confidence parameter
$\delta(\cdot)$	Dirac distribution on $\mathbb{R}^d$
δ_{jl}	Kronecker delta
i	$\sqrt{-1}$
m, n	sample size
$*_1^n$	n-fold convolution
$\mathcal{N}(\mathcal{X}, \rho, \varepsilon)$	covering number of $\mathcal{X}$
$\mathcal{H}(\mathcal{X}, \rho, \varepsilon)$	entropy number of $\mathcal{X}$
$\operatorname{span} A$	linear span of A
$\overline{x}$	complex conjuage of x
$\lvert x \rvert$	absolute value of $x \in \mathbb{C}$: $\sqrt{x\overline{x}}$

Abbreviations

Abbreviation	**Explanation**
a.e.	almost everywhere
a.s.	almost sure
MMD	maximum mean discrepancy
i.i.d.	independent and identically distributed
IPM	integral probability metric
KL	Kullback-Leibler
LCH	locally compact Hausdorff
r.h.s.	right hand side
r.k.	reproducing kernel
RKBS	reproducing kernel Banach space
RKHS	reproducing kernel Hilbert space
pd	positive definite
s.i.p.	semi-inner-product
VC	Vapnik-Červonenkis
w.r.t.	with respect to